DOI: 10.5748/19CONTECSI/PSE/DSC/7049

DATA MINING AND ANALYSIS APPLIED TO PUBLIC SECURITY DATA IN BELÉM OF PARÁ, BRAZIL

Samara Lima de Souza ; https://orcid.org/0000-0001-9807-2595
Universidade Federal do Pará

Helder Mateus dos Reis Matos ; https://orcid.org/0000-0002-5632-7948
Universidade Federal do Pará

Reginaldo Cordeiro dos Santos Filho ; https://orcid.org/0000-0002-0456-8547
Universidade Federal do Pará

João Crisóstomo Weyl Albuquerque Costa ; https://orcid.org/0000-0003-4482-6886
Universidade Federal do Pará

DATA MINING AND ANALYSIS APPLIED TO PUBLIC SECURITY DATA IN BELÉM OF PARÁ, BRAZIL

ABSTRACT

Crime is a common social problem faced around the world, and it can affect a nation's quality of life, economic growth, and reputation. Thus, law enforcement officials need to take preventive measures and one of the methods that have been gaining ground in crime analysis is data mining. With this, the purpose of this paper is to apply data analysis and data mining techniques in public security databases in the city of Belém of Pará, in order to discover hidden patterns and assist security managers in developing new public policies to try to reduce crime rates, through data from police reports in the city of Belém of Pará, Brazil in the years 2019, 2020 and 2021. To guide this study, the CRISP-DM methodology was used, where it was possible through these techniques to extract knowledge to understand and analyze the crime scene in the municipality of Belém, such as the fact that a certain crime occurs at night implies that its nature is robbery in order to assist the responsible bodies in investigations and strategies for a more effective fight against crime.

Keywords: Crime, Data Mining, Data Analytics, Association Rules, Exploratory Data Analysis.

1. Introduction

The complexity of decision-making in public security has been growing, which is natural given the various aspects that influence decisions in this area, ranging from aspects related to political and economic issues to sociocultural ones, among others. According to the 16th Brazilian Yearbook of Public Safety, published in 2022 by the Brazilian Forum on Public Safety (FBSP, 2022), the number of intentional violent deaths has decreased in Brazil, despite the improvement, extreme violence still exists, since Brazil has 2.7% of the world's inhabitants and 20.4% of murders.

Therefore, the financial resources spent on public security have increased in an attempt to decrease high crime rates. According to Koegl and Day (2019), crime is considered expensive from almost all points of view. They cite various public and private costs that are associated with, but not limited to, crime, such as the costs of tangible and intangible losses to victims, pain and suffering, lost productivity, private security, surveillance activities, and crime prevention, plus costs related to death, abuse, policing, courts, administration of justice, etc.

The amount of data and information generated daily about crime is growing. According to Prado and Colaço Júnior (2020), a great challenge that organizations responsible for public security have been facing is how to deal with a large volume of information regarding crimes and criminals. In this sense, using data to improve public safety management, guarantee the safety of the population and prevent the occurrence of crimes contributes to a society that guarantees the fundamental rights of individuals, guaranteed in article 5, of the Federal Constitution of 1988.

Thus, Information and Communication Technology (ICT) has some resources that can help to understand patterns and trends of crimes that have occurred from historical databases of occurrences. Among these, the Knowledge Discovery in Databases (KDD) process stands out, which, according to Fayyad, Piatetsky-Shapiro, and Smyth (1996), is the entire process of preparation, selection, cleaning, incorporation of prior knowledge, and the interpretation of results.

Data mining is one of the main steps for extracting knowledge in the KDD process, as well as one of the techniques that has been gaining ground to aid strategic planning and decision-making in Public Security since this term is a subdomain of artificial intelligence that can be defined as a process that aims to generate knowledge from the data and present the findings in a comprehensive way to the user. The generation of knowledge in the context of data mining can be translated into the discovery of non-trivial patterns, relationships, and trends in data useful to the user (SCHUH; REINHART; PROTE, 2019).

Based on this, this paper aims to apply analysis and data mining techniques in databases of police reports in the city of Belém of Pará in the years 2019 to 2021, through an Exploratory Data Analysis (EDA) and generation of associative rules in order to discover information that helps in the understanding of the general panorama of the crime rate, as well as to assist the public security authorities in the decision making of new public policies.

In this way, the use of data mining can provide public safety managers with privileged information, which would be more difficult to generate without the use of this resource. Thus, using as a reference the associations made from data mining, public security agencies can make decisions aimed at the safety of the population and thus further reduce cases of crime in Belém of Pará. From the data obtained, public bodies can, for example, build awareness campaigns so that the population avoids driving vehicles in risk areas, where there are high rates of theft and robbery, for example, as well as advertisements can be carried out in various media to inform the population about crime data, seeking to prevent citizens and guide themselves. Hence, through this study, the techniques applied can be

easily adapted to be used in other states of Brazil to then contribute to other research in the fight against crime.

This paper is organized as follows. In Section 2, the related works to knowledge discovery in public security databases will be reviewed. Then, in section 3, the adopted methodology is discussed. Section 4 displays the results and discussions obtained with this research, and finally, in Section 5, the conclusions and future works of the paper will be shown.

2. Related Works

In the literature, there are studies on the analysis of crime data, as well as techniques that can help to find patterns to help guide actions to prevent and combat crime to support public security policymakers. According to Singh, Kaverappa, and Joshi (2018), data mining techniques are effective in analyzing data sets and gathering useful information in many domains. In the criminal field, data mining is receiving increased attention for uncovering underlying patterns in crime data.

In Araújo and Maciel (2018), the authors used the Apriori and FP-Growth algorithms to extract association rules that generated a predictive analysis of the data, like the fact that a crime occurs at night implies that its nature is homicide. For this, data from police reports from police stations in the city of Recife, capital of the state of Pernambuco, northeast of Brazil, from January 1 to September 30, 2017, provided by the Pernambuco Social Defense Department were used.

Based on the Data Science methodology, Prado and Colaço Júnior (2020), provided an automated model supported by an experimental process to evaluate open government data related to criminal incidents of municipalities in the southeastern state of Minas Gerais, Brazil, made available by the State Secretariat of Justice and Public Safety (SEJUSP), aiming to detect associations between cities and crimes, dangerousness rankings, patterns, and models replicable in other states through the use of association rules with the Apriori algorithm to assist the process of supporting the strategic and operational decision-making of governments and law enforcement officers in the effective fight against crime.

Ratul (2020) analyzed a crime dataset from the city of Denver, USA, for the period January 2014 to May 2019 to predict and highlight trends in occurrences that, in turn, will support law enforcement agencies and the government to discover preventive measures from forecast rates, through classification algorithms such as Random Forest, Decision Tree, K-Neighbors Classifier (KNN), Linear Discriminant Analysis (LDA), AdaBoost, ExtraTrees, and four different Ensemble models to classify 15 different crime classes. The results indicated that all algorithms achieved satisfactory accuracy above 90%, except for AdaBoost. The highlight was one of the Ensemble methods that maintained the cited accuracy for the largest number of classes used.

Regateiro (2021) proposed to evaluate crime in the Pará State of northern Brazil and Belém city, the capital of Pará state, through an exploratory quantitative study, referring to the crimes of theft, robbery, vehicle robbery, murder, robbery with homicide, and bodily injury followed by death that were extracted from the Police Occurrence Bulletins in 2017, 2018, and 2019, where 4 indices were proposed, which made it possible to identify those 7 districts that remained classified with high crime in the 3 years of study, noting that they have characteristics conducive to the development of crime. The research also generated geographic views from the indices collected.

3. Methodology

This section aims to describe, clearly and objectively, the methodological aspects used to develop the present study, where the necessary steps for the development of this work will be portrayed, based on CRISP-DM (CRoss Industry Standard Process for Data Mining), which is a popular methodology used to increase the success of DM projects (Chapman, 2000) and can be applied in several business areas, where its format allows wide iteration with its phases, ranging from planning to the identification of objectives from the perspective of understanding the business to the application of the extracted knowledge, as can be seen in Figure 1.

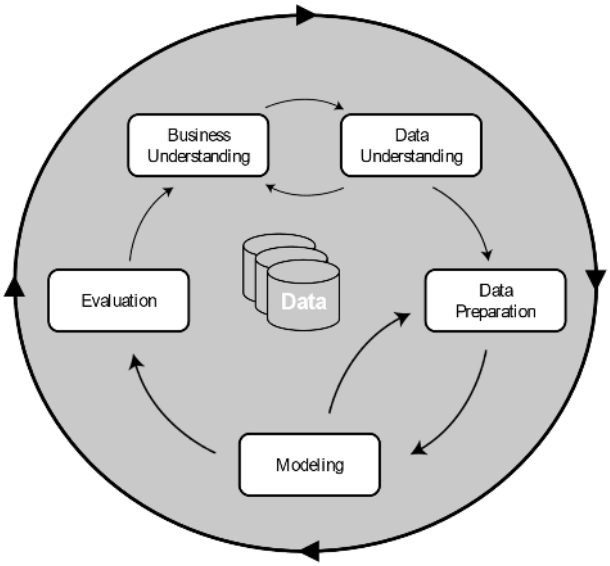

Figure. 1. Phases of the Current CRISP-DM Process Model for Data Mining.

The research was carried out in the districts of Belém city (the capital of Pará State), located in the northern region of Brazil, with an estimated population of 1,506,420 people and a territorial area of 1,059,466 km² and occupying the 11th most populous municipality in Brazil (IBGE, 2021).

3.1. Business Understanding

As seen in the previous sections, public security is a topic widely discussed in several spheres, both by citizens who suffer the direct or indirect influence caused by the insecurity feeling and by regulatory bodies of public policies, since in recent years crime has been evident as a social problem and its reduction is very important. According to ToppiReddy, Saini, and Mahajan (2018) crimes are common social problems that affect the quality of life, economic growth, and a country's reputation, as they are one of the main factors affecting several important decisions in an individual's life, such as moving to another country, a new location, avoiding risk areas.

Crimes can affect a country's economy by spending too much on fighting violence with additional police forces, providing equipment, and training, also acquiring tools to aid police work. Therefore, to minimize such problems, this study used association rules, through the Apriori algorithm. In addition to carrying out an EDA to allow an understanding of the phenomenon studied and relating variables.

3.2. Data Understanding

To achieve the goal proposed in this work, the database of police occurrences in the state of Pará, Brazil, registered in the years 2019 to 2021 was made available by SIAC, an agency linked to the Secretary of Public Security and Social Defense (SEGUP). The database can be characterized as transactional tabular data containing various information such as the crime identification, the report given by the victim describing what happened, the date and time of the occurrence, city, district, and other attributes. Counting a total of 80 attributes and 1,450,999 instances, where each instance represents a police report registered in all cities in the state of Pará. The database has records related to more than 900 different types of crimes, and nomenclatures defined by the Civil and Criminal Codes of Brazil.

3.3. Data Preparation

After understanding the provided database and the types of inputs needed to perform EDA and the application of associative rules, a preprocessing of the data was performed to obtain better results, which follows the steps presented in Figure 2.

Figure. 2. Data Preparation Steps.

Considering that the Apriori algorithm only accepts nominal variables, the first step was to select the relevant columns for rule extraction and EDA, as well as to select the instances referring to the municipality of Belém and the interest crimes.

The selected attributes were: day of the week, period of the day, the month of the event, year of the event, type of crime, means used to carry out the crime, presumed motive, districts, place of occurrence, age group of the victim, and victim's gender, where only the crimes of robbery, theft, murder, robbery with homicide, and bodily injury followed by death, chosen after meetings with the SIAC to delimit the selection of the most important crimes for this study.

In the cleaning step, when analyzing the database provided by SIAC, missing data was found in several columns. In these cases, in order not to lose information, data imputation was used, with the value "NOT INFORMED" so that the unfilled cells could be more easily detected. The value "NOT INFORMED" was added to the fields of the column's place of occurrence and the age group of the victim. The value "OTHER" was added to the fields in the presumed cause column and the value "NI" (means not informed) was added in the victim's sex column.

In the database some values like "ALREADY LAUNCHED" and "LP, PE, AND OTHERS" were found, where the first is related to a crime that had already been registered in the database, that is, it is just duplicate data, and the second refers to crime against the legal person or public entities, where for the standardization of the database these data were all removed.

Soon after, in the construction stage, the treatment of inconsistencies found in the data was done, because as the fields are open to any input, without a delimiter, some typing mistakes end up occurring, and to standardize the base, the identification and correction of

these values was done. In some fields in the database, the value "HARM" was found, indicating that the information is not known by SIAC, so to standardize the data, these values were replaced by "NOT INFORMED". And finally, the "period" column was created, containing the period in which the crime occurred: morning (06:00 - 11:59), afternoon (12:00 - 17:59), night (18:00 - 23:59), and dawn (00:00 - 05:59). Thus, after this pre-processing, the database has 174,965 rows and 11 columns.

3.4. Modeling

With the databases properly treated, an exploratory data analysis was carried out in the first place to acquire adequate knowledge about the dataset being worked on, according to Sahoo et al. (2019), EDA is an approach to see what the data can communicate, as well as helping to analyze data sets to summarize the main characteristics of attributes and how they relate to each other through information visualization techniques.

The second stage of data modeling applies to the use of data mining techniques through association rules. This a very important step, because according to Schuh et al. (2019), data mining is a subdomain of artificial intelligence that can be defined as a process that aims to generate knowledge from the data and present the findings in a comprehensive way to the user. The generation of knowledge in the context of data mining can be translated into the discovery of non-trivial patterns, relationships, and trends in data usage by the user.

Association rules are techniques used in the construction of relationships in the form of rules between items in a transactional database, seeking relationships between the attributes of objects, that is, the items that make up the base. The objective is to find strong rules according to some measure of the degree of interest in the rule (FERRARI; SILVA, 2017). According to Marzan et al. (2017), one of the best-known data mining approaches to find frequent itemsets and generate association rules is the Apriori algorithm, which was used in this study. This paved the way for researchers to solve problems on how to find hidden patterns in criminal analysis.

The application of the Apriori algorithm was used to identify rules among the variables of the database on crime, highlighting the areas of greatest occurrence, the nature of the most committed crimes, and their most frequent time, day, and month in the districts of the city of Belém. In possession of the results in the form of association rules, we have the representation of patterns of frequent relationships between items in the database studied, which allow further research and analysis regarding these rules. For this, a minimum support of 0.01 and minimum confidence of 0.7 were chosen, resulting in the generation of 106 association rules.

3.5. Evaluation

In this step, it will be presented how the process of evaluation and validation of the results obtained took place, which is a very important phase for the generation of knowledge in the databases on crime in Belém of Pará. To ensure the analysis and validation of the results of the association rules, significance tests were applied, using the support, confidence, and lift measures, to find rules that satisfy minimum significance criteria. Several support and confidence values were used to extract the best rules, where it was noticed that very high values, managed to extract few rules due to the wide variety of records in the database, and insisting on these values could cause the loss of rare and relevant itemsets. It was necessary to carry out several tests to find the ideal values as parameters to be used.

4. Results and Discussions

This section presents the results obtained through the data analysis and data mining techniques for crime fighting applied in this work, by an EDA and association rule extraction, using data from the incident reports of the years 2019, 2020, and 2021 in the municipality of Belem of Pará.

4.1. Exploratory Data Analysis

To better understand and extract knowledge from the database, an EDA was carried out through data visualization techniques, which is one of the most efficient techniques to represent and find answers with the data.

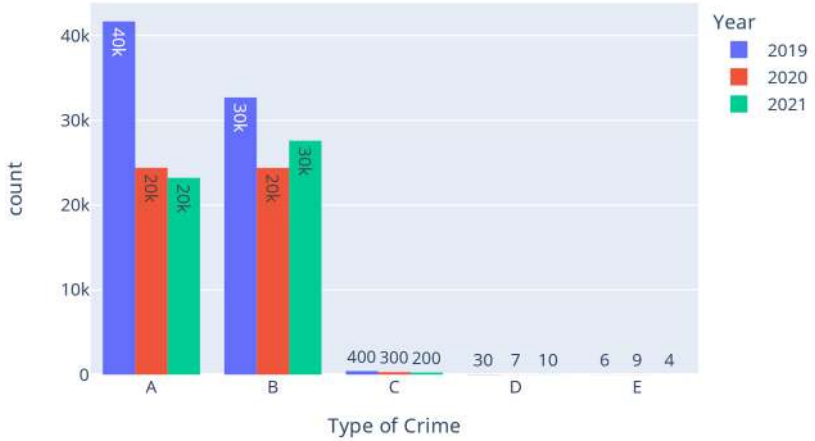

Figure 3. Frequency of Crimes per Year.

Among the crimes selected for the analysis, it can be seen in Figure 3 that the crime of theft (A) and robbery (B) were the most incidents in the municipality of Belém of Pará in the years 2019 to 2021, where more than 170,000 crimes occurred, while the less frequent crimes were murder (C), robbery with homicide (D), and bodily injury followed by death (E). In addition to noting that there was a significant decrease of approximately 41% in thefts in the year 2019 to 2020, there was also a decrease of 25% in robberies from 2019 to 2020 and an increase of 13% from 2020 to 2021.

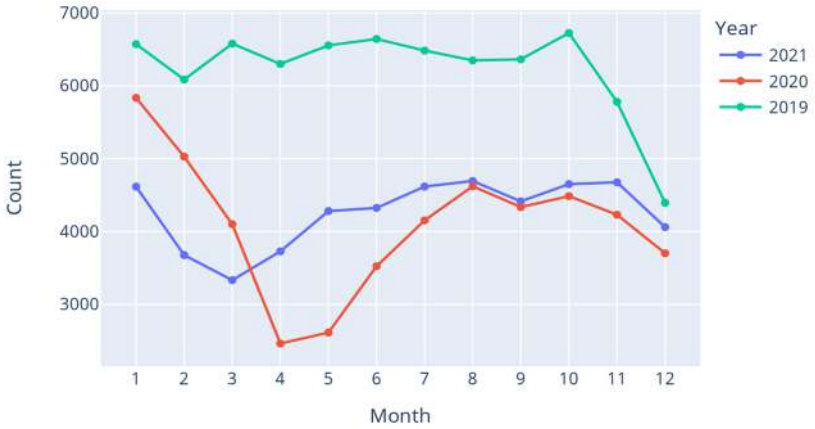

Figure 4. Frequency of Crimes over the Months.

According to the data presented in Figure 4, the crime rate of the occurrence of crimes (theft, robbery, murder, robbery with homicide, and bodily injury followed by death) over the months the years 2019 to 2021, where it is not possible to find a pattern for the records of crimes, can be explained until due to the occurrence of the 2019 coronavirus (COVID-19) pandemic in the years 2020 and 2021, wherein February 2020, according to Farias (2020), we had the first confirmed case and in April to December of the same year, the number of records of crimes were the lowest among the months of all years. In 2019, the highest incidence of crimes took place in October (10), and in 2020 and 2021 were the months of January (01) and August (08) respectively.

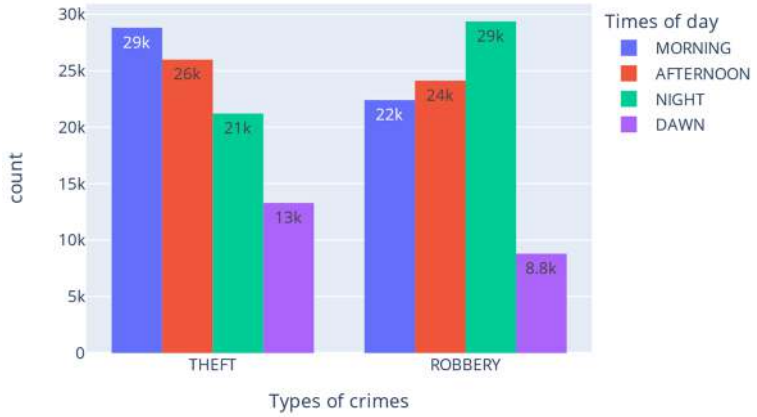

Figure 5. Periods of the day of theft and robbery crimes in the years 2019 to 2021.

When analyzing Figure 5 and Figure 6, it is possible to notice that depending on the crime, the period of the day with the highest incidence varies according to the crime being treated, as it can be seen that the crime of theft was more frequent on the part of the morning and the number of records decreased throughout the day, the crime of robbery and murder happened more frequently at night and the crimes of robbery with homicide and bodily injury followed by death (B.I.F.D) had a very similar number of records during the periods of the day.

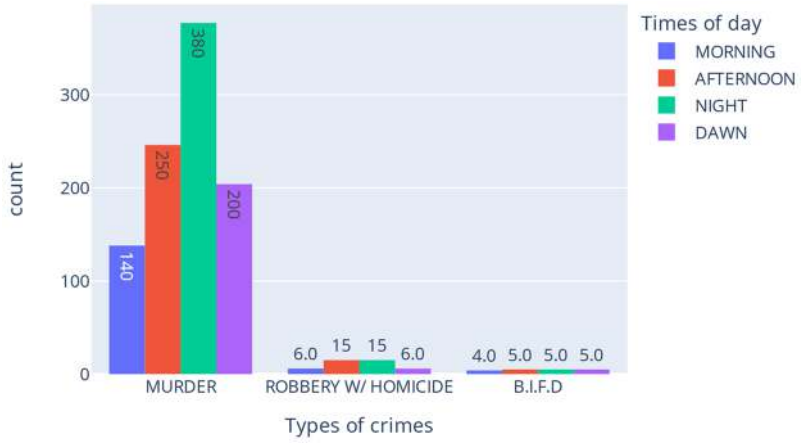

Figure 6. Periods of the day for the crimes of murder, robbery with homicide, and bodily injury followed by death in the years 2019 to 2021.

With the exposition of the data in Figure 7, it is possible to identify the 10 districts that had the most incidences of crimes (theft, robbery, murder, robbery with homicide, and bodily injury followed by death) in the years 2019 to 2021, where the Marco district was the

most frequent, with 11,302 occurrences, in second place the Guamá district with 10,367, in third the São Brás district, and so on.

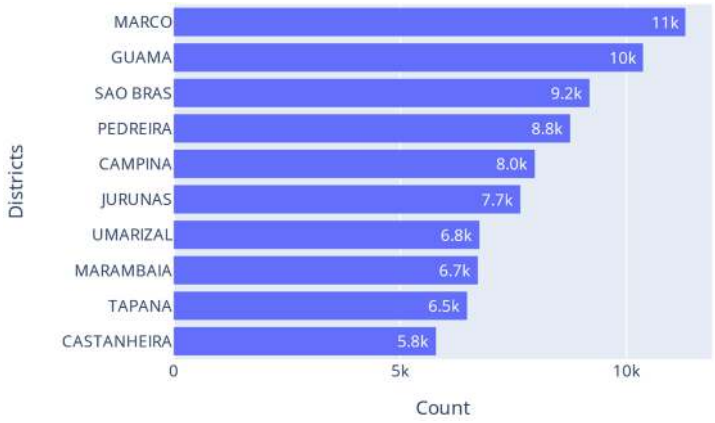

Figure 7. Most Frequent Districts.

In Figure 8, it is possible to analyze that most victims of crimes (theft, robbery, murder, robbery with homicide, and bodily injury followed by death) in the years 2019 to 2021 are adults aged between 35 and 64 years old and male, as well as in all age groups the predominant sex is male, whereas the elderly are in the age group of 65 years or older, adult IV (35-64 years), adult III (30-34 years), adult II (25-29 years), adult I (18-24 years), adolescent (12-17 years), and child (0-11 years old).

Figure 8. Age Pyramid of Victims of Crimes.

Hence, by this analysis, it is possible to understand the crime scene in the city of Belém of Pará because knowing the most frequent crime, the period of the day when each crime occurs, the month, district, age, and gender most frequent is possible alert the community and assist the responsible bodies in the creation of public policies.

4.2. Association Rules

After the EDA, the Apriori algorithm was applied to extract association rules relevant to the analysis in order to correlate the attributes present in the database. Thus, the post-processing database was used, containing 174,965 records and the minimum support value was set as 0.01, and the minimum confidence value was set as 0.7, which generated a total of 106 rules, where only 6 were highlighted for performing a more detailed analysis, presented below:

***Rule 1: IF** the sex is MALE, the means used is WITHOUT INSTRUMENT, the year of the fact is 2019, and the place of occurrence is VIA PUBLIC **THEN** the type of crime is THEFT. Sup = 0.05, Conf = 86%, Lift = 1.7.*

***Rule 2: IF** the district is MARCO, the victim's age group is ADULT IV (35 TO 64 YEARS), and the year of the fact is 2019 **THEN** the crime type is THEFT. Sup = 0.02, Conf = 75%, Lift = 1.5.*

***Rule 3: IF** the victim's age group is ADULT IV (35 TO 64 YEARS), the month of the fact is OCTOBER, and the year of the fact is 2019 **THEN** the crime type is THEFT. Sup = 0.01, Conf = 79%, Lift = 1.6.*

***Rule 4: IF** the means used is WITHOUT INSTRUMENT, the victim's age group is ADULT IV (35 TO 64 YEARS), and the period is MORNING **THEN** the crime type is THEFT. Sup = 0.06, Conf = 97%, Lift = 1.9.*

***Rule 5: IF** the means used is FIREARM and the period is NIGHT **THEN** the type of crime is ROBBERY. Sup = 0.08, Conf = 97%, Lift = 2.0.*

***Rule 6: IF** the means used is FIREARM, the sex of the victim is MALE, and the age group of the victim is ADULT IV (35 TO 64 YEARS) **THEN** the type of crime is ROBBERY. Sup = 0.05, Conf = 97%, Lift = 1.9.*

After analyzing the rules, we can perceive and understand some characteristics of each crime, given rules 1, 2, 3, and 4 that deal with a theft crime, some conclusions and strategies for preventing crime can be drawn, as they portray that the perpetrator of theft most of the time does not use a physical instrument to carry out the crime, in addition to his most frequent victim being male and adult between 35 and 64 years old, during the morning period (06:00 – 11:59), in a public road, in the Marco district, in October, and more specifically in 2019.

Thus, some strategies for police prevention and policing can be taken, such as, for example, police patrols should be deployed on the streets between 06:00 and 11:59, when a greater proportion of theft crimes occur in the Marco district, on public roads, to make more strategic decisions.

Rules 5 and 6 deal with the crime of robbery, where some differences from the crime analyzed above can be seen, since the robbery, unlike the crime of theft, the most used means is the firearm and the night period (18: 00 – 23:59) is the most frequent. In this way, the police must be aware of suspicious behavior during this period in men aged 35 to 64 years, particularly among those who may be carrying a firearm, which is very important information that allows concluding the need to build more operational policies more effective in monitoring and controlling the movement of firearms.

It is important to emphasize that the association rules generated with the attributes present in the database do not mean a direct causal connection, but that there is a high correlation between them and that they should be analyzed by a public security specialist in order to obtain the best results and assist in the creation of new police strategies.

5. Conclusions and Future Work

Crime is a problem that affects everyone and can cause great damage to the economy and social growth of a country. For this reason, this paper presented the application of data mining and analysis techniques to help in the fight against crime, using data extracted from police reports for the years 2019, 2020, and 2021 in Belém of Pará, Brazil, made available by SIAC.

The techniques used in this study could benefit the police, law enforcement authorities, and public security in decision-making and the creation of new public policies, through exploratory data analysis and data mining techniques, such as the extraction of association rules in order to find patterns in databases.

Through the results obtained with the exploratory data analysis, it was possible to have a clearer understanding of the database used, finding hidden patterns and non-trivial information, by way of information visualization techniques to facilitate the comprehension of the generated data, such as knowing the distribution of criminal occurrences over the months of the years worked and realizing that 2019 was the year with the most occurrences in all months.

In addition to the previous technique, the Apriori algorithm of association rules was used, resulting in the extraction of patterns of frequent relationships between items from the database used, allowing for in-depth research and analysis of the generated rules. The results showed that the predictions generated had an average confidence value greater than 70%, confirming the relevance of the information, making it possible, through them, for the police to act more effectively in the fight and prevention of crimes, as they will know, for example, which periods of the day will need to intensify the police fleets on the streets, which are the characteristics of the crime, as well as identify which districts need higher priority in the investigations.

Thus, the main contribution of this study was to demonstrate that analysis and data mining techniques can be used together to help obtain knowledge about public security in the studied region and consequently in other regions in order to help in the fight against crime.

In future work, we intend to use other databases to try to justify crime spikes, trying to relate socioeconomic data with crime, for example. Furthermore, there is the potential use of other machine learning techniques to try to find different results and explore the possibilities of analysis through science and data mining.

References

Araújo, B. C., & Maciel, A. M. A. (2018). Aplicação de regras de associação em dados da criminalidade da cidade do Recife. *Revista de Engenharia e Pesquisa Aplicada*, *3*(3).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. Wirth, R. (2000). CRISP-DM 1.0 - Step-by-step data mining guide, *CRISP-DM Consortium*. 76(3).

Farias, H. S. D. (2020). O avanço da Covid-19 e o isolamento social como estratégia para redução da vulnerabilidade. *Espaço e Economia. Revista brasileira de geografia econômica*, (17).

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

FBSP. (2022). Anuário brasileiro de segurança pública. *Brasil*.

Ferrari, D. G., Silva, L. N. D. C. (2017). *Introdução a mineração de dados*. Saraiva Educação SA.

IBGE. (2021). Instituto brasileiro de geografia e estatística.

Koegl, C. J., Day, D. M. (2019). The monetary costs of crime for a sample of offenders in Ontario. *Canadian Journal of Criminology and Criminal Justice*, 61(3):21– 44.

Marzan, C. S., Baculo, M. J. C., de Dios Bulos, R., Ruiz Jr, C. (2017). Time series analysis and crime pattern forecasting of city crime data. In *Proceedings of the International Conference on Algorithms, Computing and Systems*, pages 113–118.

Prado, K. H. J, Colaço Júnior, M. (2020). Data science aplicada á análise criminal baseada nos dados abertos governamentais de minas gerais. *Research, Society and Development*, 9(11), e36391110044-e36391110044.

Ratul, M. A. R. (2020). A comparative study on crime in Denver city based on machine learning and data mining. *arXiv preprint arXiv:2001.02802*.

Regateiro, H. A. S. (2021). Avaliação da criminalidade em belém e no estado do Pará. page 296.

Sahoo, K., Samal, A. K., Pramanik, J., Pani, S. K. (2019). Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(12):2019.

Schuh, G., Reinhart, G., Prote, J.-P., Sauermann, F., Horsthofer, J., Oppolzer, F., Knoll, D. (2019). Data mining definitions and applications for the management of production complexity. *Procedia CIRP*, 81:874–879.

Singh, N., Kaverappa, C. B., Joshi, J. D. (2018). Data mining for prevention of crimes. In *International Conference on Human Interface and the Management of Information*, pages 705–717. Springer.

ToppiReddy, H. K. R., Saini, B., Mahajan, G. (2018). Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science*, 132:696–705.