Data Mining in Public Security Databases in Belém, Pará, Brazil

Samara Lima de Souza[1], Helder Mateus dos Reis Matos[1], Cleyton Fernando Paixão de Sousa Costa[1], Reginaldo Cordeiro dos Santos Filho[1], João Crisóstomo Weyl Albuquerque Costa[1]

[1]Universidade Federal do Pará, R. Augusto Correa, Guamá, Belém, 66075-110, Pará, Brasil.

{samara.souza, helder.matos}@icen.ufpa.br, cleyton.costa@ifch.ufpa.br, {regicsf, jweyl}@ufpa.br

Abstract. *Crime is a common social problem faced worldwide that can affect a nation's quality of life and economic growth. 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 the development of new public policies to try to reduce crime rates. Through this study, it was possible to obtain results that can help public security authorities to understand crime, as well as in making decisions about new security policies.*

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 issues, as well as economic, and sociocultural, 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.

According to [Singh et al. 2018], data mining techniques have proven effective in analyzing datasets and gathering useful information in many domains. In the criminal field, data mining is receiving increased attention to discover underlying patterns in crime data, as can be seen in [Prado et al. 2020, Araújo and Maciel 2018, Ratul 2020, Regateiro 2021].

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 association 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.

2. Methodology

The methodology used was 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 et al. 2000]. The first stage of CRISP-DM consists of understanding

the business, where it was possible to identify that public security is a theme widely discussed in various spheres, both by citizens who suffer the direct or indirect influence caused by the feeling of insecurity and by the regulatory bodies of public policies since in recent years crime has been evident as a social problem and its reduction is important.

The second step consisted in understanding the data, where the database used was made available by the Assistant Secretariat for Intelligence and Criminal Analysis (SIAC), an agency linked to the Secretariat of Public Security and Social Defense (SEGUP). The database contains records of police occurrence reports recorded in the years 2019 to 2021, containing various information such as the crime identification, the date and time of the occurrence and registration, city, district, and other attributes. Counting a total of 80 attributes and 1,450,999 instances. The database has records related to more than 900 different types of crimes.

The data preparation step had three phases. The first phase was the data selection, where 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, districts, place of occurrence, victim's age group, and victim's gender, where only 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 crimes.

In the second phase, data cleaning was performed, where missing values found in some columns were treated with the imputation technique, assigning values such as "not informed" in the detected rows. In the third phase, the treatment of inconsistencies found in the data was carried out, since some values had typing errors, as well as 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 had 175,965 rows and 14 columns.

In the modeling step, an EDA was carried out on the data, to acquire knowledge about the dataset that is being worked on, through information visualization techniques. For the generation of visualizations, the programming language, Python, was used, widely used in science and data mining tasks. The second stage of data modeling employs association rule techniques through the application of the Apriori algorithm, which was used to identify rules among the features of the crime database. For this, a minimum support of 0.01 and minimum confidence of 0.7 were chosen, which resulted in the generation of 106 association rules.

In the evaluation step, 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.

3. Results and Discussions

3.1. 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. Among the crimes selected for the analysis, it can be seen in Figure 1 that the crime of theft (A) and robbery (B) were the most

incidents in the city 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.

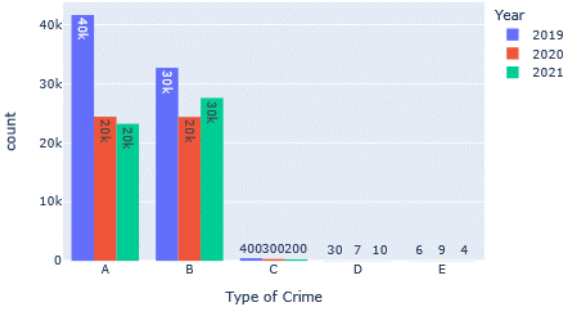

Figure 1. Frequency of crimes per year

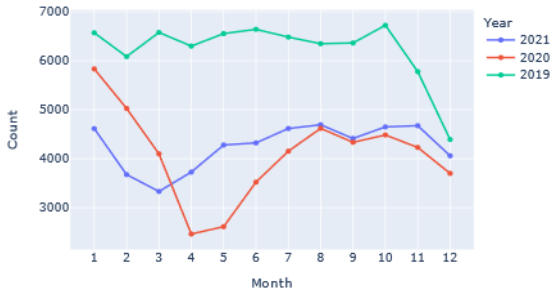

Figure 2. Frequency of crimes over the Months

According to the data presented in Figure 2, the crime rate of the occurrence of crimes 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 we had the first confirmed case in Brazil [Farias 2020] 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, and in 2020 and 2021 the months of January and August respectively.

3.2. Association Rules

After the EDA, the Apriori algorithm was applied to extract associative rules relevant to the analysis, in order to correlate the features, present in the database. Thus, the post-processing database was used, which generated a total of 106 rules, where only 4 were highlighted for performing a more detailed analysis, presented below:

Rule 1: 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 2: 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 3: 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 4: 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 and 2 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, in the Marco district, 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 to make more strategic decisions.

Rules 3 and 4 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 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.

5. Conclusion

The results presented in this paper have shown that the use of data mining and analysis techniques can benefit areas such as public safety, since, through the insights obtained with EDA, it was possible to have a clearer understanding of the database worked, finding hidden patterns and non-trivial information, through information visualization

In addition to the previous technique, the Apriori algorithm of associative rules was used, resulting in the extraction of patterns of frequent relationships between items from the database, where 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.

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.

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).

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

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.

Ratul, M. A. R. (2020). A comparative study on crime in Denver city based on machine learning and data mining. CoRR, abs/2001.02802.

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

Singh, N., Bellathanda Kaverappa, C., Joshi, J. D. (2018). Data mining for prevention of crimes. In International Conference on Human Interface and the Management of Information (pp. 705-717). Springer, Cham.