Research, Society and Development, v. 10, n. 13, e177101320744, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i13.20744

Mineração de dados e análise de ocorrências criminais: um estudo sobre o Município de Divinópolis – Minas Gerais

Data mining and analysis of criminal occurrences: a study on the Municipality of Divinópolis – Minas Gerais

Minería de datos y análisis de ocurrencias delictivas: un estudio sobre el Municipio de Divinópolis – Minas Gerais

Recebido: 17/09/2021 | Revisado: 25/09/2021 | Aceito: 05/10/2021 | Publicado: 09/10/2021

Felipe Augusto Souza Mamedes
ORCID: https://orcid.org/0000-0002-2625-5269
Universidade do Estado de Minas Gerais, Brasil
E-mail: felipe.1694494@discente.uemg.br
Marcos Alberto Saldanha
ORCID: https://orcid.org/0000-0001-7218-9883
Universidade do Estado de Minas Gerais, Brasil
E-mail: marcos.saldanha@uemg.br
Edwaldo Soares Rodrigues
ORCID: https://orcid.org/0000-0002-1546-5500
Universidade do Estado de Minas Gerais, Brasil
E-mail: edwaldo.rodrigues@uemg.br

Resumo
A criminalidade é um problema que os governos e sociedades enfrentam, sendo realizados diversos investimentos em segurança e inteligência pública, para tentar punir e prevenir ações criminais. Esse projeto tem como objetivo auxiliar na segurança pública, aplicando técnicas de Mineração de Dados sobre bases de dados contendo boletins de ocorrências criminais da cidade de Divinópolis-MG. As bases de dados contam com ocorrências de janeiro de 2016 a maio de 2019, providas pela Polícia Militar do Estado de Minas Gerais (PMMG). Como procedimentos metodológicos foi inicialmente realizada a seleção dos dados, em seguida, o pré-processamento e transformação dos dados. Na sequência, aplicou-se técnicas de mineração de dados, tais como: Clusterização e Extração de Regras de Associação. Além disso, dedicou-se uma etapa na qual foram realizadas análises estatísticas relacionadas aos crimes de "Furto" e "Roubo", bem como crimes relacionados à violência contra a mulher. Dentre os resultados destacam-se duas Regras de Associação, encontradas com o uso do algoritmo Apriori, "Noite, Roubo" => "Vítima do sexo Masculino" e **"**Armas de Fogo" => "Sem lesão aparente", além das regras foram realizadas análises estatísticas sobre os dados, como "análise das faixas etárias das vítimas" e "distribuição das ocorrências criminais na semana". Dessa forma, conclui-se que, este trabalho chegou aos objetivos almejados, trazendo conhecimentos que podem ser utilizados pelos órgãos de segurança pública. Finalmente, sugere-se como trabalhos futuros, a ampliação da base de dados, bem como trabalhar com dados de latitude e longitude para cada ocorrência criminal.
Palavras-chave: Análise criminal; Mineração de dados em ocorrências criminais; Extração de regras de associação.

Abstract
Crime is a problem that governments and societies face, with several investments being made in public security and intelligence, to try to punish and prevent criminal actions. This project aims to assist in public safety, applying data mining techniques on databases containing criminal incident bulletins in the city of Divinópolis-MG. The databases have occurrences from January 2016 to May 2019, provided by the Military Police of the State of Minas Gerais (PMMG). The methodological procedures, data selection was initially performed, followed by data pre-processing and transformation. Next, data mining techniques were applied, such as: Clustering and Extraction of Association Rules. In addition, the stage was dedicated to statistical analysis related to crimes of "Theft" and "Robbery", as well as crimes related to violence against women. Among the results, two Association Rules stand out, found using the Apriori algorithm, "Night, Robbery" => "Male Victim" and "Firearms" => "No apparent injury", in addition to the statistical rules were performed on the data, such as "analysis of the age groups of victims" and "distribution of criminal occurrences in the week". Thus, it is concluded that this work reached the desired goals, bringing knowledge that can be used by public security agencies. Finally, it is suggested as future works, the expansion of the database, as well as working with latitude and longitude data for each criminal occurrence.
Keywords: Criminal analysis; Data mining in criminal occurrences; Extraction of association rules.

Resumen
La delincuencia es un problema que enfrentan los gobiernos y las sociedades, con diversas inversiones en seguridad pública e inteligencia para tratar de sancionar y prevenir acciones delictivas. Este proyecto tiene como objetivo ayudar en la seguridad ciudadana, aplicando técnicas de Data Mining en bases de datos que contienen boletines de incidentes delictivos en la ciudad de Divinópolis-MG. Las bases de datos tienen ocurrencias de enero de 2016 a mayo de 2019, provistas por la Policía Militar del Estado de Minas Gerais (PMMG). Como procedimientos metodológicos, inicialmente se realizó la selección de datos, seguida del preprocesamiento y transformación de los datos. A continuación, se aplicaron técnicas de minería de datos, tales como: Clustering y Extracción de Reglas de Asociación. Además, se dedicó una etapa a los análisis estadísticos relacionados con los delitos de "Hurto" y "Robo", así como los delitos relacionados con la violencia contra la mujer. Entre los resultados destacan dos Reglas de Asociación, encontradas mediante el algoritmo Apriori, "Noche, Robo" => "Víctima Masculina" y "Armas de Fuego" => "Sin lesión aparente", además de las reglas donde se realizaron análisis estadísticos sobre los datos, como "análisis de los grupos de edad de las víctimas" y "distribución de los hechos delictivos en la semana". Así, se concluye que este trabajo alcanzó las metas deseadas, aportando conocimientos que pueden ser utilizados por los organismos de seguridad pública. Finalmente, se sugiere como trabajos futuros, la ampliación de la base de datos, así como trabajar con datos de latitud y longitud para cada ocurrencia delictiva.
Palabras clave: Análisis criminal; Minería de datos en sucesos delictivos; Extracción de reglas de asociación.

1. Introdução

Devido ao crescimento demográfico de algumas regiões, junto à crise econômica e alta nas taxas de desemprego, a criminalidade se tornou um problema ainda mais recorrente. Como diz Dornelles (2017), o problema da violência criminal é apenas consequência de vários outros problemas sociais de um país.

O resultado de um crime não reflete apenas em suas vítimas diretas, mas para todo o país. Alguns dos reflexos no Brasil são exemplificados por Diniz (2005), como: altos gastos com segurança pública e visitantes internacionais abaixo de seu potencial turístico. Minas Gerais, por mais que tenha taxas de criminalidade menores quando comparado aos outros estados do país, também se encaixa na situação supracitada. Scalco (2007) mostra que o estado apresentou um crescimento de 459% nas taxas de crimes violentos entre 1986 e 2005.

O município de Divinópolis pertence a Região Centro-Oeste de Minas Gerais. Segundo dados do Censo 2020 – IBGE (Instituto Brasileiro de Geografia e Estatística) a cidade tem uma área territorial de 708,115 km² e uma população estimada de 240.480 habitantes. O município não foge a regra de seu estado. De acordo com notícia publicada pelo portal G1 Centro-Oeste (2019)[1], em agosto de 2019 Divinópolis ocupava a 13ª posição entre as cidades mineiras com maior número de homicídios registrados, a pesquisa foi realizada pelo Instituto de Pesquisa Econômica Aplicada (Ipea) expondo altas taxas criminais.

Nesse contexto, são gerados vários dados sobre as ocorrências criminais, como data, horário, local, tipo de crime e informações dos envolvidos. *Big Data*, como define Machado (2018), é caracterizada por conjuntos de dados volumosos e que crescem constantemente. Os dados gerados pelas ocorrências criminais podem ser considerados como conjuntos de *Big Data*, devido suas características. Hammond (2013) diz que a partir desse alto volume de dados é possível gerar evidências, e essas por sua vez, podem apresentar soluções para diversos problemas.

Considerando o aumento contínuo da capacidade de armazenamento, cada vez mais dados são preservados. Sendo assim, cria-se a possibilidade de aplicação de algumas técnicas de Mineração de Dados, onde se pode extrair padrões ou identificar registros que são considerados pontos fora da curva, comumente denominados *outliers*, gerando assim, informações preciosas para as organizações e sociedade. de Amo (2004) afirma que, técnicas de Mineração de Dados começaram a ser utilizadas no início dos anos 80, quando as organizações perceberam que estavam acumulando muitos dados, mas em contrapartida não os utilizavam. Ainda de acordo com a autora supracitada, Mineração de Dados pode ser definido de forma simples, sendo a extração ou mineração de conhecimento em conjuntos de dados volumosos.

[1]https://g1.globo.com/mg/centro-oeste/noticia/2019/08/06/atlas-da-violencia-divinopolis-ocupa-a-13a-posicao-entre-as-cidades-mineiras-com-a-maior-taxa-de-homicidios.ghtml

De acordo com de Amo (2004) as técnicas de Mineração de Dados se aprimoraram, a autora afirma que nos primórdios da Mineração de Dados, esta era usada apenas para extrair uma informação das bases de dados volumosas da forma mais rápida possível. No entanto, nos dias atuais são utilizadas diversas técnicas de Mineração de Dados, e além disso, uma das principais etapas, consiste na análise dos dados, que são gerados pelas técnicas utilizadas, possibilitando assim o descobrimento de conhecimento que sejam mais relevantes, e de modo a agregar valor. Além disso, pode-se evidenciar que novos algoritmos foram desenvolvidos, fator que também contribui para a área de um modo geral. No cenário atual, as técnicas são utilizadas por áreas diversas, como *marketing* para analisar o comportamento do consumidor e direcionar anúncios e ofertas personalizadas, bancos para verificar as transações do usuário e identificar possíveis fraudes, dentre outras, inclusive na área criminal conforme neste trabalho e outros que serão apresentados no Capítulo 3.

Apesar das grandes bases de dados, um estudo que gere evidências relevantes é trabalhoso e inviável se não computacionalmente. Partindo disso a proposta desse artigo é analisar uma base de dados fornecida pela Polícia Militar da cidade de Divinópolis – MG. Foram realizados os passos de pré-processamento, realizadas análises estatísticas e utilizadas técnicas de mineração de dados, como Extração de Regras de Associação e Clusterização, com intuito de gerar conhecimento relacionado à segurança pública do município de Divinópolis-MG.

Visto isso, para realização desse projeto, o restante do artigo foi dividido da seguinte maneira. No Capítulo 2 há o referencial teórico, com algumas definições para que o leitor possa ter um melhor entendimento do trabalho. No Capítulo 3 são apresentadas obras relacionadas ao tema trabalhado. No Capítulo 4 realiza-se uma descrição sobre a base de dados utilizada. No Capítulo 5 é discorrido sobre a metodologia aplicada. No Capítulo 6 são expostos os resultados obtidos. Por fim, no Capítulo 7 tem-se a conclusão deste trabalho, bem como sugestões de trabalhos futuros.

2. Referencial Teórico

Neste capítulo serão apresentados alguns termos para melhor entendimento do artigo. Na Seção 2.1, é discutido sobre a obtenção da base de dados utilizada. Na Seção 2.2, são apresentados alguns conceitos de Regras de Associação. Na Seção 2.3, são exibidas algumas definições referentes à Clusterização. Por fim, na Seção 2.4, serão expostos estudos e pesquisas sobre violência contra a mulher.

2.1 Sobre as bases de dados

Uma das dificuldades para realizar Mineração de Dados sobre bases de dados criminais é a obtenção dos dados. Para obtê-los, pode-se realizar uma mineração em portais de notícias por temas criminais. Contudo, dessa forma, além de ser um processo trabalhoso, acaba gerando bases de dados incompletas, pois não há um formulário de preenchimento para os dados relacionados as ocorrências criminais nos portais de notícias, faltando dados que seriam essenciais na aplicação de técnicas de Mineração de Dados. Outro ponto a ser discutido neste cenário é a ausência de padronização entre as notícias, o que dificulta no processo de análises realizados pelos algoritmos de Mineração de Dados.

Neste sentido, visando eliminar as dificuldades mencionadas anteriormente, para este trabalho foi estabelecido contato diretamente com a Polícia Militar de Minas Gerais (PMMG), do município de Divinópolis-MG, onde foi apresentada a ideia de se aplicar técnicas de Mineração de Dados com o intuito de se descobrir conhecimento na base de dados de ocorrência criminal, com o objetivo de auxiliar inclusive a PMMG na tomada de decisões estratégicas.

Mediante este contato realizado, houve a autorização da PMMG para o compartilhamento da base de dados de ocorrências criminais do município de Divinópolis, havendo-se a exigência de se respeitar o termo de privacidade para fins acadêmicos, bem como impossibilitando o compartilhamento de dados isolados e ou individualizados. Em posse da base de dados, verificou-se que a base de dados disponibilizada contava inicialmente com dois arquivos, no formato '.xlsx', totalizando

171.100 registros contendo boletins de ocorrências de janeiro de 2016 a dezembro de 2017 e de janeiro de 2018 a maio de 2019. A base de dados será melhor descrita no Capítulo 4.

2.2 Extração de Regras de Associação

A Extração de Regras de Associação é uma estratégia que busca por padrões frequentes e que estão de certa forma escondidos na base de dados.

As regras costumam ser expressas como **X → Y,** onde se lê: X implica em Y. No exemplo X é o antecedente e Y o consequente da regra, e significa que SE ocorre X na base de dados ENTÃO espera-se que ocorra também Y.

Visto isso, existem ainda algumas métricas que avaliam o quão pertinente é a regra de associação. Dentre as métricas mais importantes, destacam-se:

- **Suporte:** representa o percentual de vezes que a regra aparece na base de dados. Como exemplo de interpretação de uma regra, se o suporte de uma regra for 0,05 ou 5%, significa que esta regra se repete em 5% da base de dados. Desta forma, quanto maior o valor do suporte, maior a representatividade da regra na base dados. Contudo, o fato de mais representativa, não significa elucidar um novo conhecimento, talvez essa maior representatividade indique algo que já é conhecido e não agregue nada de importante. A Equação 1 a seguir apresenta a definição do suporte.

$$\text{Sup}(A \rightarrow B) = \frac{\text{Registros que contêm } A \rightarrow B}{\text{Registros totais}} \quad (1)$$

- **Confiança:** representa o quão confiável é a regra. Quando a regra tem 80% ou 0,8 de confiança, quer dizer que 80% das vezes que o antecedente aparece, ocorre o consequente também. Ao passo que o valor para a métrica confiança aumenta, mais confiável será a regra de associação. A Equação 2 a seguir apresenta a definição do suporte.

$$\text{Conf}(A \rightarrow B) = \frac{\text{Registros que contêm } A \rightarrow B}{\text{Registros que contêm } A} \quad (2)$$

A seguir, na Subseção 2.2.1, será apresentado o Apriori, o algoritmo utilizado para realizar a Extração de Regras de Associação.

2.2.1 Apriori

Dentre os algoritmos que trabalham com Extração de Regras de Associação se destaca o algoritmo Apriori. É mostrado no trabalho de Prado et al. (2020) que o Apriori é o mais utilizado quando se tratando de Regras de Associação.

Neste trabalho também optou-se por utilizar o algoritmo Apriori devido seu fácil entendimento e alta performance. De acordo com Marzan et al. (2017) e Romão et al. (1999), o algoritmo Apriori é uma das abordagens mais conhecidas e atraentes para gerar regras de associação. Existem duas funções no Apriori segundo Romão et al. (1999), sendo a primeira utilizada na identificação e separação dos itens que são frequentes, já a segunda é responsável pela extração das regras de associação identificadas no conjunto de itens frequentes, gerados pela primeira função.

2.3 Clusterização

Clusterização, como definido por Ochi et al. (2004), consiste em dividir os elementos quaisquer em grupos, nos quais objetos semelhantes fiquem no mesmo *cluster* e objetos diferentes fiquem em *clusters* separados.

Os algoritmos utilizados podem ser divididos em dois subgrupos:

1) **Supervisionados:** são algoritmos que recebem nos parâmetros tanto os valores de entrada quanto os de saída. Comumente, realiza-se antes um treinamento com os dados rotulados, para posteriormente o algoritmo poder rotular outras entradas.
2) **Não-supervisionados:** nesse caso o objetivo também é separar conjuntos de dados em grupos, contudo as saídas não são conhecidas. O algoritmo mesmo sem rótulos anteriores verifica os atributos e padrões que se repetem dentre os dados e faz a divisão dos grupos.

Existem diversas metodologias e algoritmos para clusterização, seguindo diferentes passos para definir os *clusters*, a seguir, na Subseção 2.3.1, serão apresentados o algoritmo K-means e, na Subseção 2.3.2, o Método Elbow, utilizados neste trabalho.

2.3.1 K-means

Os resultados da pesquisa de Prado et al. (2020) demonstram que o K-means é o algoritmo mais utilizado nos métodos não-supervisionados.

Visto isso, o K-means foi o algoritmo selecionado para a clusterização neste trabalho. Trata-se de um método não supervisionado pois não há grupos pré-definidos na base de dados trabalhada, o algoritmo tem que encontrar as semelhanças entre os objetos e fazer a separação sozinho. O funcionamento do K-means pode ser descrito da seguinte forma:

1. O K-means inicialmente coloca K centroides em pontos aleatórios.
2. Cada item próximo passa a pertencer ao grupo do centroide mais próximo.
3. O reposicionamento dos centroides é calculado fazendo a média das distâncias de cada elemento do grupo.
4. Os passos 2 e 3 são repetidos até que os *Clusters* fiquem estáveis.

Ao fim das iterações são gerados K *clusters* distintos de acordo com o parâmetro passado.

2.3.2 Método Elbow

A escolha da quantidade K de *clusters* foi realizada com auxílio do método Elbow, uma técnica que busca o valor ideal de grupos para o processo de clusterização. Segundo Kodinariya & Makwana (2013), o método do cotovelo (traduzindo para o português) é um dos mais clássicos para determinar a quantidade de *clusters* de um conjunto de dados, além de ser um método visual. O método Elbow consiste em iniciar com K igual a 2 e realizar os cálculos da variança dos *clusters*, aumentando a cada passo o K em 1. Em um certo momento a variância terá uma queda e estabilizará, mesmo que aumente o K. No ponto que ocorrer essa queda é a quantia de *clusters* ideal.

2.4 Violência contra a mulher

Os diversos crimes que ocorrem contra as mulheres apenas por serem mulheres são resumidos pelo termo 'violência contra a mulher'. Esses crimes não são apenas aqueles que resultam em lesões físicas, mas também lesões psicológicas, morais, dentre outras. Segundo o portal G1 (2019)[2], no ano de 2019 aumentaram em 7,3% os casos de feminicídios no Brasil, isto é, crimes que são realizados por causa do gênero da vítima.

Segundo Engel (2020), nas últimas duas décadas ocorreu uma melhora relevante com relação a coleta de dados relacionados ao problema de violência contra a mulher. Da mesma forma, estão melhores utilizados os dados, seja pelo Estado brasileiro para pensar em soluções, ou pelas organizações feministas para realizarem cobranças de ações do governo.

[2]https://g1.globo.com/monitor-da-violencia/noticia/2020/03/05/mesmo-com-queda-recorde-de-mortes-de-mulheres-brasil-tem-alta-no-numero-de-feminicidios-em-2019.ghtml

Entretanto, ainda de acordo com a autora, pode-se dizer que a violência contra a mulher nos últimos dez anos aumentou, no entanto, talvez isso ocorra justamente por causa do aumento de dados pesquisados, devido a isso, ainda não é possível afirmar estatisticamente com precisão se houve uma diminuição ou aumento nas taxas desse tipo de crime no Estado brasileiro.

A violência contra a mulher é um problema de saúde pública, que ocorre independente da classe socioeconômica, no Brasil e vem se tornando cada vez mais evidente gerando uma necessidade das autoridades e sociedade trabalharem para combatê-la. A melhora na coleta de dados permite realizar estratégias mais precisas para soluções do problema, focando em regiões e tipos de crimes que estão mais críticos.

Por sua vez neste trabalho foi realizada uma análise na base de dados buscando padrões e regras com foco nos crimes contra a mulher, que será apresentada na Seção 5.5.

3. Trabalhos Relacionados

Tayal et al. (2015) propõe implementar técnicas de Mineração de Dados para identificar crimes em cidades indianas. A base de dados foi extraida a partir de conjuntos de dados não estruturados de várias fontes de crimes na *Web*, do período de 2000 – 2012. Os autores fizeram comparações dentre as fomas utilizadas pelas organizações de detecção de crimes e os métodos propostos por eles. Para aprimorar os resultados, os autores utilizaram o algoritmo K-means incorporando os mapas do Google, visando identificar as regiões dos criminosos. Aplicaram também o K Nearest Neighbor (KNN) com o intuito de analisar se um criminoso que tenha cometido um crime no passado, pode ou não ser o autor de um crime atual. Ou se a ocorrencia atual associa-se a um padrão dos crimes já registrados anteriormente.

É apresentada por Pereira e Brandão (2014) a ARCA (Association Rules for Crime Analysis) uma abordagem para encontrar regras em conjuntos de dados criminais. Os dados utilizados foram fornecidos pelo governo brasileiro, mediante termo de privacidade, contendo dois anos de ocorrências criminais reais. O método utilizado no ARCA foi o algoritmo Apriori. Com esse trabalho foi possível observar os picos de horários dos crimes de furto, entre 8h e 11h, e roubo, 20h. Foi identificado também a possível causa de crimes violentos realizados por jovens, em sua maioria motivados por questões financeiras. Por fim, diagnosticaram que os crimes de roubo a mão armada não resultam em ferimentos.

Os autores Sevri, Karacan e Akcayol (2017) utilizaram a Mineração de Dados sobre um conjunto de dados de crimes reais dos EUA de 2013 registrados pelo FBI contendo aproximadamente 5 milhões de dados. A metodologia abordada foi o uso do algoritmo Apriori para geração de regras de associação, onde foi configurado um suporte de 0,05 e confiança de 0,6. Este trabalho gerou várias regras, dentre elas, "se as raças da vítima e do agressor forem brancas e a cena do crime for em casa, então o sexo da vítima é masculino".

Já Marzan et al. (2017), propõe identificar as áreas de alta criminalidade a partir do conjunto de dados analisado. A fonte de dados utilizada foi coletada manualmente no Escritório do Distrito Policial de Manila (MPD), contando inicialmente com dados criminais dos anos de 2012 a 2016 de 16 distritos. A estratégia aplicada foi KDE (Kernel Density Estimation) para verificar a densidade de crimes na cidade, além da utilização do algoritmo Apriori com intuito de gerar regras de associação. Por meio do KDE foi possível visualizar graficamente 3 regiões com maior incidência criminal. O Apriori gerou diversas correlações entre tipo de crime e região, com mais regras para as 3 regiões que se destacaram anteriormente. Não foi gerada nenhuma regra para crime de Abuso Sexual, sugerindo que esse crime ocorra em situações aleatórias.

Por fim, há o estudo de Prado e Júnior (2020), que utiliza bases de dados contendo informações das cidades de Minas Gerais dos anos de 2012 à 2020, disponíveis devido à política de transparência pública no Brasil. Os autores aplicam técnicas de Mineração de Dados para extrair Regras de Associação. No total são encontradas 136 regras de associação com o padrão "Município => Tipos de Crime". É realizado também um *ranking* de Perigosidade.

Engel (2020) faz uma análise sobre Violência contra a Mulher no Brasil. A autora utiliza em seu trabalho dados públicos de pesquisas relacionados ao tema. Mediante a análise dos dados verifica-se que, a maior parte dos crimes de violência contra mulher são cometidos por homens conhecidos, mesmo quando fora de casa. Além disso, nota-se que, o local o qual ocorrem mais agressões de mulheres é em casa. Também foi percebido que, mulheres jovens e negras sofrem mais violências seja dentro ou fora de casa. Ainda são apresentadas outras estatísticas como tipo de lesões, tipo de denúncia, dentre outras.

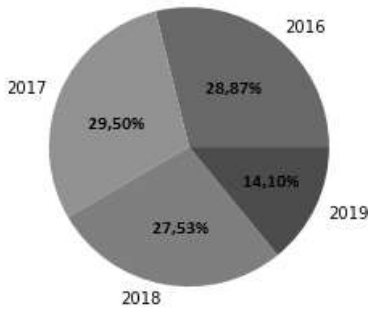
Por mais que neste artigo e nos anteriores o objeto de estudo seja a criminalidade e o uso de *data science* para tentar reduzir tal problema, há diferenças. A principal diferença é a base de dados, tanto pelo tamanho quanto as informações contidas. Outra distinção é a forma de trabalhar com a base de dados, já que este trabalho, além de realizar uma análise para a base toda, realizou também outras duas análises uma com foco nos crimes de roubo e furto e outra com foco nos crimes os quais a vítima é do sexo feminino.

4. Descrição da Base de Dados

Como já citado anteriormente, a base de dados foi adquirida com a Polícia Militar de Minas Gerais (PMMG) do município de Divinópolis-MG. Respeitando o termo de privacidade para fins acadêmicos, não pode-se compartilhar registros isolados, contudo, pode-se demonstrar análises e estudos realizados sobre os dados.

A base inicialmente era formada por dois arquivos no formato '.xlsx', o primeiro continha dados de Boletins de Ocorrências criminais dos anos de 2016 e 2017 e o segundo continha dados de 2018 e de 2019 até o mês de julho. Todos os dados totalizam 171.100 registros distribuídos de acordo com o gráfico da Figura 1 a seguir.

Figura 1. Representação gráfica da distribuição de registros por ano.



Fonte: Desenvolvido pelos autores.

No gráfico da Figura 1 nota-se que, 49.393 registros são do ano de 2016, 50.479 registros são referentes ao ano de 2017, 47.101 registros de 2018 e por fim 24.127 registros referentes aos meses de janeiro a julho de 2019.

Cada registro na base de dados original é descrito por 30 atributos, conforme apresentado na Tabela 1.

Tabela 1. Base original, atributos que descrevem os registros e suas descrições.

Nome do Atributo	Descrição	Nome do Atributo	Descrição
Número REDS	Código de identificação para aocorrência	Tipo Envolvimento	Autor // Vítima // Dentre outros
Qtde. Envolvidos	Na maioria dos registros valor 1	Sexo	Sexo do indivíduo do registro
Data Comunicação do Fato	Dia, mês e ano da que foi registrada aocorrência	Cútis	Cútis do indivíduo
Bairro	Bairro no qual aconteceu a ocorrência	Embriaguez	Nível de embriaguez do indivíduo
Logradouro Ocorrência	Endereço do ocorrido	Turista	Sim // Não
Logradouro Ocorrência – Tipo	Tipo de endereço: Rua // Avenida //Rodovia	Ocupação Atual	Trabalho do indivíduo
Descrição Subclasse Nat. Principal	O que foi a ocorrência	Estado Civil	Estado Civil do indivíduo
Tentado/Consumado Nat. Principal	Tentado // Consumado	Escolaridade	Escolaridade do indivíduo
Descrição Subgrupo Compl. Nat.	O que está envolvido na ocorrência	Idade Aparente	Idade aparente do indivíduo
Latitude	-	Nacionalidade	Nacionalidade do indivíduo
Longitude	-	Município Naturalidade	Município Natural do indivíduo
Data Fato	Dia, mês e ano da ocorrência	UF Naturalidade – Sigla	UF do indivíduo
Horário Fato	Horário da ocorrência	Grau Lesão	Grau das lesões do indivíduo quando se aplica
Qtde. Envolvidos	Coluna duplicada	Causa Presumida	Possível causa que gerou a ocorrência
Relação Vítima/Autor	-	Desc. Longa MeioUtilizado	Meios utilizados pelos infratores

Fonte: Desenvolvido pelos autores.

Apesar de inicialmente a base de dados conter 30 atributos, nem todos estes foram utilizados no desenvolvimento deste projeto, uma vez que alguns, não possuem uma relevância para o propósito do presente trabalho. No Capítulo 5 serão apresentados e discutidos melhor os atributos utilizados.

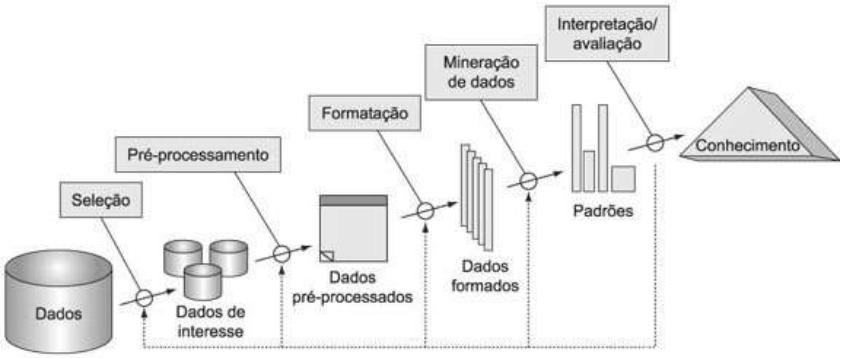
5. Metodologia

A metodologia é uma das principais etapas em um trabalho acadêmico, onde os métodos científicos são apresentados pelo conjunto de procedimentos realizados, que possuem por intuito prezar pelo conhecimento em um trabalho científico, conforme destacado por Prodanov e De Freitas (2013). O método científico abordado no presente trabalho possuiu natureza básica, onde a forma de abordagem do problema foi quantitativa. Em relação aos fins da pesquisa, essa foi explicativa, tendo como procedimentos a combinação entre a pesquisa de campo e a *ex-post-facto*.

Há na literatura diversos estudos que abordam a temática de Mineração de Dados e suas técnicas, sendo inclusive estudos de alta relevância para essa área de estudos. Neste sentido, é importante ressaltar que não há motivos para criar uma metodologia do início ao fim, sendo que já tem estudos que trazem importantes conceitos. Desta forma, este projeto

inicialmente fará uso da abordagem metodológica de Fayyad et al. (1996), conforme apresenta-se na Figura 2, que divide a metodologia de execução de projetos de Mineração de Dados em 5 etapas, compreendidas desde a seleção dos atributos até a interpretação e análise dos *insights* identificados.

Figura 2. Etapas para Mineração de Dados.



Fonte: Adaptado de Fayyad et al. (1996)

Apesar de ser utilizada a abordagem ilustrada na Figura 2, é importante ressaltar que, o projeto aqui desenvolvido possui suas especificidades. Sendo assim, fará uso da metodologia de modo a alinhar com as características e especificidades da base de dados, podendo assim não apresentar exatamente o mesmo fluxo ou exatamente os mesmos passos abordados em Fayyad et al. (1996).

Para este trabalho utilizou-se a linguagem de programação Python, principalmente devido aos recursos que esta linguagem oferece, tais como bibliotecas diversas e específicas sobre temas os quais foram abordados no trabalho. Sendo assim, foram utilizadas bibliotecas como Pandas, utilizada no processo de limpeza e pré-processamento dos dados; Scikit-learn, aplicada no momento da elaboração e execução dos algoritmos de Mineração de dados, entre outras.

Como apresentado na Figura 2, a primeira etapa é a seleção dos dados que será desenvolvida na Subseção 5.1. As etapas seguintes são as de pré-processamento e depois transformação dos dados que serão trabalhadas na Subseção 5.2. Por fim verificam-se as etapas de Mineração de Dados e a de Interpretação dos resultados, que serão apresentadas na Subseção 5.3. Além dessas etapas, serão também descritos, na Subseção 5.4, o processo de Análise sobre os Crimes de Furtos e Roubos e, na Subseção 5.5, a Análise dos Crimes Contra a Mulher.

5.1 Seleção dos Dados

Em grandes bases de dados nem sempre todos os registros e atributos são úteis, sendo necessária uma seleção dos conjuntos de dados importantes e remoção dos inválidos e irrelevantes antes de trabalhar com a base. Esse processo de seleção é importante, inclusive, para que os algoritmos aplicados consigam gerar melhores resultados, uma vez que atributos com as características supracitadas podem prejudicar o desempenho do algoritmo.

Conforme apresentado na Tabela 1, inicialmente cada registro da base de dados era descrito por 30 atributos. Após verificar o conteúdo e qual o tipo de informação que este trazia, foram descartados aqueles duplicados e aqueles que na maioria dos registros possuíam valores nulos, inválidos ou ausentes. Alguns dos atributos descartados foram: Causa Presumida, Embriaguez, Estado Civil, Relação Vítima/Autor, dentre outros, que estavam em muitos registros sem preenchimento ou preenchidos como “OPCIONAL”, “IGNORADO”. Ressalta-se que esses tipos de preenchimentos referem-se a atributos que o policial ao preencher o boletim de ocorrência não teria a obrigatoriedade de preencher.

Desta forma, os atributos que foram selecionados passaram pela etapa de pré-processamento conforme será descrito na subseção seguinte.

5.2 Pré-processamento e Transformação dos Dados

Os registros inicialmente se encontram inapropriados para serem trabalhados, sendo necessárias transformações e preparações antes de realizar o processo de mineração.

Nesta etapa foi verificado que quando o atributo 'Número REDS' era o mesmo em diferentes registros, isso significava que ambos registros se referiam a uma mesma ocorrência, sendo assim, foi realizada a união de ocorrências equivalentes em um único registro, gerando assim uma nova base de dados, agora com 91.405 registros. Ao realizar a união foi considerado apenas 1 autor e 1 envolvido/vítima, preenchendo esse campo como "Não Def" quando a ocorrência não havia um deles.

Além disso, foi realizada a limpeza na base de dados, removendo alguns registros com muitos atributos nulos e também uma padronização dos valores que não passavam informações como aqueles preenchidos por 'OPCIONAL'. Ainda, foram aplicadas as seguintes *features enginers:* por meio do atributo 'Data', foram adicionados os atributos 'DiaSemana', 'Feriado' e 'FimSemana', que serão descritos na Tabela 2 posteriormente, por meio dos atributos 'AutorIdade' e 'EnvolvidoIdade', foram adicionados os atributos 'AutorFaixaEtaria' e 'EnvolvidoFaixaEtaria'. Na literatura não há uma divisão clara sobre as faixas etárias, então foram adotadas as seguintes: Criança (0-12), Adolescente (13-19), Jovem (20-29), Adulto (30-39), Adulto2 (40-59) e Idoso (60+).

Ao fim deste processo, a base de dados passou a ter 24 atributos, conforme se verifica na Tabela 2.

Tabela 2. Base preprocessada, atributos que descrevem os registros e suas descrições.

Nome do Atributo	Descrição	Nome do Atributo	Descrição
CodOcorrencia	Código de identificação para a ocorrência	Causa	Possível causa que gerou a ocorrência
Bairro	Bairro no qual aconteceu a ocorrência	Meio	Meios utilizados pelos infratores
CrimeTipo	O que foi a ocorrência	AutorSexo	Sexo do autor
Latitude	-	AutorCutis	Cor da pele do autor
Longitude	-	AutorEscolaridade	Nível de escolaridade do autor
Hora	Horário que aconteceu a ocorrência	AutorIdade	Idade do autor
Ano	Ano que foi registrada a ocorrência	AutorFaixaEtaria	Faixa etária (de acordo com classificação supracitada) do autor
DiaSemana	Qual dia da semana aconteceu a ocorrência	EnvolvidoSexo	Sexo da vítima
Feriado	Se a ocorrência aconteceu em um feriado ou não	EnvolvidoCutis	Cor da pele da vítima
FimSemana	Se a ocorrência aconteceu em um final de semana ou não	EnvolvidoEscolaridade	Nível de escolaridade da vítima
Turno	Madrugada, manhã, tarde ou noite	EnvolvidoIdade	Idade da vítima
Lesão	Grau das lesões do indivíduo do registro quando se aplica	EnvolvidoFaixaEtaria	Faixa etária da vítima

Fonte: Desenvolvido pelos autores.

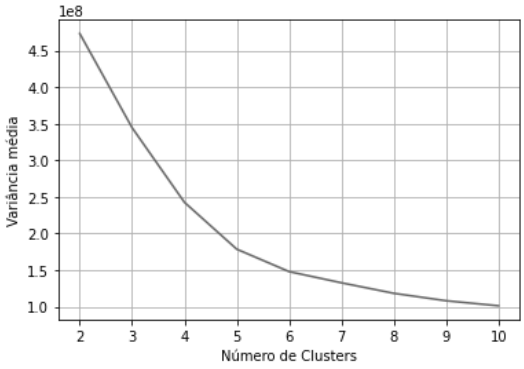
5.3 Mineração de Dados e Interpretação dos Resultados

Este passo consiste na aplicação dos algoritmos para buscar padrões e regras na base de dados. Para este trabalho dividiu-se a etapa de Mineração de Dados entre o processo de Clusterização e Extração de Regras de Associação nos *Clusters*, que será descrito na Subseção 5.3.1, e a Extração de Regras de Associação em toda a base de dados, sendo esta exposta na Subseção 5.3.2.

5.3.1 Clusterização e Extração de Regras de Associação nos *Clusters*

Inicialmente a base de dados passou por um processo de clusterização, isso foi realizado com objetivo de encontrar regras mais específicas e ocultas do conjunto ao aplicar o algoritmo Apriori sobre os grupos separados. Ressalta-se que o ponto inicial para a realização do processo de clusterização é a definição da quantidade de *clusters*, para isso, este trabalho se utilizou do método de Elbow, conforme se verifica na Figura 3.

Figura 3. Gráfico Elbow.



Fonte: Desenvolvido pelos autores.

No gráfico da Figura 3, o eixo X representa a quantia de c*lusters* enquanto o eixo Y representa a variância. Nota-se por meio do gráfico que, a quantia de *clusters* ideal fica dentre 5 ou 6 *clusters*, visto que após isso, mesmo se há um aumento nos *clusters* não há diminuição significativa na variância.

Após utilizar o método Elbow para decidir a quantidade de *clusters*, foi realizada por meio do K-means a clusterização em 5 grupos, número este que está entre a quantidade ideal de *clusters*. Salienta-se que após a clusterização cada *cluster* agrupou os tipos de crimes identificados na base de dados. Ressalta-se ainda, que como este algoritmo é não supervisionado, o agrupamento pode conter dados que ao olho nu, não se mostram de certa forma similares.

Sobre cada um dos *clusters* foi aplicado o algoritmo Apriori, utilizando diferentes parâmetros de suporte e confiança, começando com as métricas mais altas 0,8 de confiança e 0,1 de suporte e reaplicando o algoritmo reduzindo as métricas a cada aplicação, chegando até 0,5 de confiança e 0,01 de suporte.

5.3.2 Extração de Regras de Associação em toda a Base de Dados

Conforme mencionado na seção anterior, o algoritmo Apriori foi utilizado com o intuito de se realizar a Extração de Regras de Associação em cada um dos *clusters*, visando encontrar regras de acordo com itens que estivessem em um mesmo *cluster*, já que em tese, estes itens possuem mais semelhanças. Contudo, de modo a Extrair Regras de Associação em toda a base de dados, foi também executado o algoritmo Apriori para todo o conjunto de dados.

A princípio, o algoritmo foi aplicado, com valores de suporte e confiança mais altos, respectivamente 0,8 e 0,1, buscando regras gerais da base. Contudo, esses valores mais altos não trouxeram conhecimentos relevantes, e então, foram realizadas novas aplicações com menores métricas utilizadas para o Apriori, diminuindo o valor da confiança até 0,5 e suporte até 0,01. Vale ressaltar que, a cada execução, foi realizada uma análise sobre as regras para verificar se estas traziam algum conhecimento.

Por fim, foram realizadas outras aplicações do Apriori, mas agora em busca de regras filtradas, considerando se dentre os antecedentes havia certo atributo e/ou se dentre os consequentes havia certo atributo. Nessa etapa, os valores de suporte e confiança foram aplicados da mesma forma que na etapa anterior, inicialmente foram utilizados os valores mais altos reduzindo-os a cada aplicação até chegar em valores menores para cada uma das métricas, diminuindo o valor da confiança até 0,5 e suporte até 0,02, a cada aplicação verificando se dentre as regras obtidas alguma trazia conhecimento.

5.4 Análise sobre os Crimes de Furtos e Roubos na Base de Dados

Devido o fato de crimes de furtos e roubos terem a maior quantidade de ocorrências na base de dados, estes foram selecionados para uma análise mais específica, uma vez que são responsáveis por afetar a sociedade de forma mais rotineira. O intuito desta análise foi identificar padrões nesses crimes, seja nas características das vítimas, horários, ou bairros com maior incidência destes crimes, visto que essas ocorrências constituem aproximadamente 25% da base de dados.

Os crimes de roubo e furto representam, respectivamente, aproximadamente 15,8% e 6,8% do total das 91.405 ocorrências da base de dados, o que consiste em aproximadamente 25% de toda a base de dados. Vale ressaltar que, furto e roubo são crimes diferentes de acordo com o Código Penal Brasileiro[3], onde, furto consiste em: subtrair o patrimônio de outra pessoa sem violência; já roubo consiste em: um crime mais grave caracterizado pela subtração do patrimônio alheio mediante violência ou grave ameaça à pessoa.

Para realizar a análise sobre essa parte dos dados, foi inicialmente criada uma outra base de dados filtrando dos registros principais apenas os crimes de “Furto” e “Roubo”. Depois, na nova base de dados gerada, com 20.721 registros, foi aplicado o algoritmo Apriori para realizar a Extração das Regras de Associação que foram posteriormente analisadas. Além disso também foram realizadas análises estatísticas.

5.5 Análise dos Crimes Contra a Mulher

Conforme citado na Subseção 2.4, os crimes de violência contra a mulher são um problema de saúde pública que necessita de atenção. Diante desta situação, e com o intuito de tentar identificar conhecimento na base de dados sobre essa situação tão difícil e que afeta milhares de mulheres no dia a dia, este trabalho buscou contribuir nesta temática, uma vez que trazer esse assunto e eventualmente algum conhecimento sobre o mesmo pode ser um pequeno passo para as autoridades darem uma maior atenção, além de pensar possibilidades de mitigar essa questão.

Inicialmente, nesta etapa realizou-se uma filtragem na base de dados, de modo a gerar uma nova base de dados, contendo apenas as ocorrências as quais a vítima era do sexo feminino. Após realizar esse processo, foi obtida uma nova base de dados com 31.840 registros, representando 34,8% dos registros totais. Sobre esses dados filtrados foram realizadas análises estatísticas e aplicação do algoritmo Apriori visando a Extração de Regras de Associação que pudessem trazer conhecimento acerca do tema abordado.

[3]http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm

6. Resultados

Neste capítulo serão apresentados os resultados obtidos em cada etapa e os conhecimentos gerados por eles. Na Seção 6.1, é discutido sobre os resultados relacionados às Regras de Associação nos *clusters*. Na Seção 6.2, são apresentados os resultados da Extração de Regras de Associação em toda a base de dados. Na Seção 6.3, são exibidas as conclusões geradas a partir das análises sobre os crimes de furtos e roubos. Por fim, na Seção 6.4, serão expostas as análises e conclusões que se deram a partir dos crimes contra a mulher.

6.1 Extração de Regras de Associação nos *Clusters*

Conforme apresentado na Seção 5.1, foram realizadas inicialmente aplicações do algoritmo Apriori em *clusters* com o intuito de encontrar regras que estivessem relacionadas mais intimamente a um determinado grupo de dados. A cada aplicação, variando as métricas de "suporte" e "confiança", foram geradas centenas de regras, com diferentes valores de antecedentes e consequentes, essas foram analisadas uma a uma para verificar se traziam algum conhecimento.

A Tabela 3, a seguir, apresenta algumas das regras extraídas nessa etapa com o algoritmo Apriori.

Tabela 3. Regras extraídas dos *Clusters.*

Cluster	Regra		Métricas	
	Antecedentes	**Consequentes**	**Suporte**	**Confiança**
5	'Fraude', 'Vantagem Econômica', ' Vítima Sexo Feminino'	'Sem Lesões Aparentes', 'Estelionato'	0,040	0,977
3	'Autor Sexo Masculino', 'Sem Lesões Aparentes', 'Agressão Física Sem Emprego De Instrumentos'	'Vias De Fato / Agressão'	0,041	0,617
2	'Briga / Atrito', ' Vítima Sexo Masculino'	'Ameaça'	0,046	0,862
5	'Falta De Atenção'	'Abandono Do Local De Acidente De Trânsito', 'Sem Lesões Aparentes', 'Veiculo'	0,047	0,803
1	'Armas De Fogo', ' Vítima Escolaridade Alfabetizado', 'Vantagem Econômica'	'Roubo', ' Vítima Sexo Masculino'	0,048	0,731
2	'Estelionato'	'Sem Lesões Aparentes', 'Vantagem Econômica'	0,049	0,876
3	'Atrito Familiar'	'Agressão Física Sem Emprego De Instrumentos'	0,054	0,652
1	'Roubo', ' Vítima Sexo Masculino', 'Noite'	'Armas De Fogo', 'Sem Lesões Aparentes', 'Vantagem Econômica'	0,059	0,711
4	'Roubo', 'Vantagem Econômica', 'Noite'	'Armas De Fogo'	0,083	0,831
4	'Vítima Pele Branca', 'Armas De Fogo', 'Sem Lesões Aparentes'	'Roubo'	0,085	0,931

Fonte: Desenvolvido pelos autores.

Dentre as regras geradas, conforme visto na Tabela 3, a maioria não trazia um conhecimento novo ou era redundante. Contudo destacaram-se as Regras de Associação que tinham como ideia principal "Armas de Fogo" => "Sem lesão aparente", encontradas nos *clusters* 2, 4 e 5. Essa regra sugere que as armas de fogo são utilizadas na maioria das vezes apenas como instrumentos de intimidação no ato do crime não gerando lesões físicas.

6.2 Extração de Regras de Associação em toda a base de dados

Para a busca de regras na base de dados integral foi inicialmente apenas aplicado o Apriori e analisadas as regras geradas. O objetivo dessa etapa foi encontrar regras mais genéricas, padrões que se apliquem para toda a base.

Após a busca mais genérica foi aplicado o Apriori buscando regras filtradas. Nessa etapa filtrou-se as regras por aquelas que tivessem certos atributos nos antecedentes ou consequentes, foram aplicados filtros em busca de diferentes atributos, com intuito de encontrar algo mais específico que estivesse oculto na base de dados.

Da mesma forma que foi encontrado em alguns grupos da etapa anterior, nesta etapa também foi encontrada a regra "Armas de Fogo" => "Sem lesão aparente", variando com alguns antecedentes e/ou consequentes a mais, com valores de "suporte" e "confiança" de até aproximadamente 0,04 e 0,6, conforme mostrado na Tabela 4 a seguir.

Tabela 4. Regras extraídas da base integral.

Regra		**Métricas**	
Antecedentes	**Consequentes**	**Suporte**	**Confiança**
'Sem Lesões Aparentes', 'Armas De Fogo'	'Roubo', 'Vítima Sexo Masculino'	0,042	0,638
'Roubo', 'Vítima Sexo Masculino'	'Sem Lesões Aparentes', 'Armas De Fogo'	0,042	0,659
'Vantagem Econômica', 'Armas De Fogo'	'Sem Lesões Aparentes', 'Roubo'	0,042	0,775
'Sem Lesões Aparentes', 'Vantagem Econômica', 'Armas De Fogo'	'Roubo'	0,042	0,961
'Vantagem Econômica', 'Armas De Fogo'	'Roubo'	0,052	0,962

Fonte: Desenvolvido pelos autores.

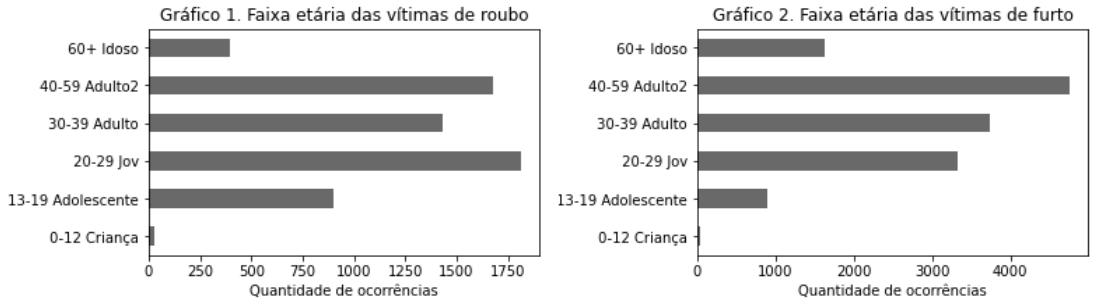
Verifica-se na tabela que mesmo com as variações de antecedentes e consequentes a principal ideia é "Armas de Fogo" => "Sem lesão aparente". Além dessa regra, também foi encontrada outra regra que pode trazer informações interessantes, como: "Noite, Roubo" => "Vítima do sexo Masculino" com o valor da métrica "confiança" de aproximadamente 0,7 e "suporte" de 0,03. Talvez esse padrão ocorra devido o fato de os homens se sentirem mais seguros de andar sozinhos a noite na rua, e exatamente por isso são as vítimas mais frequentes nesse tipo de crime e neste horário.

6.3 Análise sobre os crimes de Furto e Roubo

Após as extrações de regras de toda a base de dados, tentou-se identificar conhecimento relacionado aos crimes de "Furto" e "Roubo". Primeiramente executou-se o algoritmo Apriori filtrando apenas as regras relacionadas aos crimes escolhidos para a análise. Dentre as regras encontradas nenhuma apresentava algum conhecimento aparente.

Então foram realizadas análises estatísticas, onde inicialmente foram verificadas quais eram as faixas etárias das vítimas dos crimes de "Roubo" e "Furto" conforme se verifica nos gráficos da Figura 4.

Figura 4. Análise das faixas etárias das vítimas.

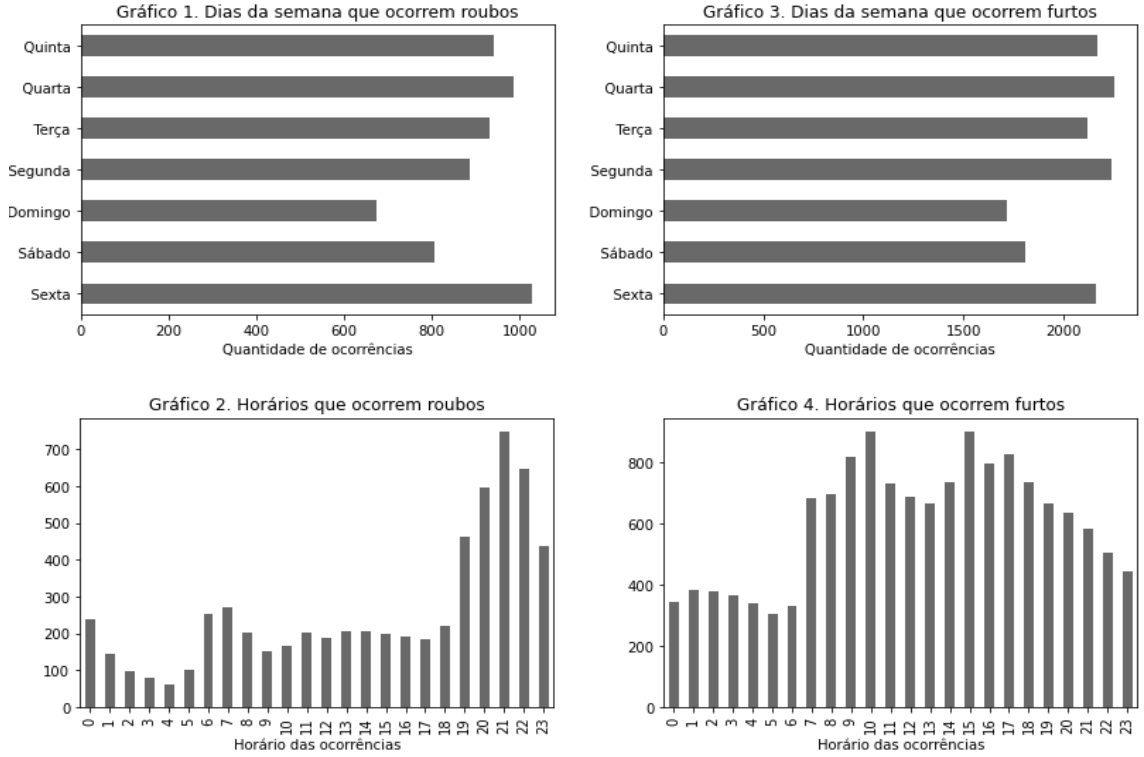


Fonte: Desenvolvido pelos autores.

Nota-se na Figura 4 que para ambos os tipos de crime, “Furto” e “Roubo”, na maioria das ocorrências as vítimas envolvidas estão entre 20 e 59 anos. Há um fluxo muito maior de pessoas circulando nas cidades nessa faixa, seja a trabalho ou lazer, e talvez isso justifique a maior concentração dessas idades como vítimas de crimes, representando mais de 75% do total.

Foi realizada também uma análise com relação aos horários e dias em que ocorrem os crimes conforme os gráficos da Figura 5.

Figura 5. Gráficos relacionados aos crimes de “Furto” e “Roubo”.



Fonte: Desenvolvido pelos autores.

Verifica-se nos Gráficos 1 e 3 da Figura 5 que não há uma diferença discrepante entre os dias da semana que ocorrem ambos os tipos de crimes. Contudo com relação aos horários, nota-se no Gráfico 2, que os crimes de “Roubo” ocorrem em maior número a noite, possivelmente por serem crimes contra alvos isolados, então aproveitam-se da noite pois é o turno com menor fluxo de pessoas e sem a luz do dia, facilitando a execução do crime.

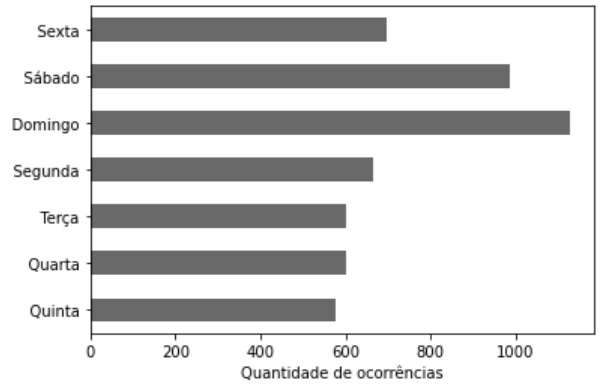
Já com relação aos crimes de “Furto”, no Gráfico 4, verifica-se que ocorrem em maior número durante a tarde, o que provavelmente se justifica porque são crimes que ocorrem, ao contrário dos crimes de “Roubo”, enquanto as ruas estão cheias, aproveitando a multidão para extrair algo de um cidadão distraído ou invadir as residências vazias enquanto os residentes estão a trabalho ou estudo.

6.4 Análise dos Crimes Contra a Mulher

Com relação aos crimes contra a mulher foi inicialmente aplicado o algoritmo Apriori em uma base de dados derivada da base principal, a qual foram filtrados aqueles crimes cujo a vítima era do sexo feminino. As Regras de Associação geradas pelo algoritmo nessa base de dados não trouxeram nenhum conhecimento que possa ser considerado, já que a maioria das regras apresentavam teor já conhecido, ou pouco relevantes para a análise. Então foram realizadas análises estatísticas.

Foi realizada uma busca aos tipos de crimes que mais ocorriam contra a mulher, nessa busca destacou-se os crimes de 'vias de fato / agressão' e 'lesão corporal', somando por volta de 15% do total dos crimes contra mulher. Na base de dados principal, mais de 50% dos crimes de lesão corporal foram contra mulheres e mais de 65% dos crimes de agressão também foram contra vítimas do sexo feminino. Na Figura 6 a seguir, verifica-se a distribuição das ocorrências relacionadas a esses crimes durante a semana.

Figura 6. Crimes de agressão e lesão contra a mulher, distribuição de ocorrências na semana.

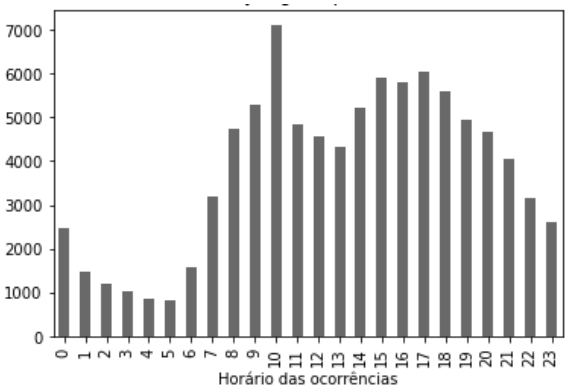


Fonte: Desenvolvido pelos autores.

Percebe-se, ao analisar o gráfico da Figura 6 que, a maioria das ocorrências de agressão e lesão contra mulher são registradas durante os fins de semanas, sábado e domingo. Isso pode corroborar o estudo apresentado por Engel (2020), que ressalta que a maior parte das ocorrências de violência contra mulher são efetuadas por homens conhecidos, e exatamente aos fins de semana são quando geralmente os indivíduos passam a maior parte do tempo juntos em suas residências.

Durantes as análises, foram verificados picos em diferentes tipos de crimes contra mulher às 10 horas. Foram realizadas diversas buscas com o Apriori utilizando como filtro se no antecedente ou consequente havia o atributo de horas igual a 10h. Também foram plotados gráficos diversos em busca de algo relacionado a este atributo que se destacou. Contudo, não foi encontrado nenhum conhecimento aparente que justifique a causa dos picos de crime às 10 horas. A Figura 7 a seguir mostra esse fato de forma gráfica.

Figura 7. Crimes contra a mulher, distribuição de geral por horários.



Fonte: Desenvolvido pelos autores.

7. Conclusão

Conforme apresentado anteriormente, a criminalidade é um problema social que assombra diversas questões da vida pública. Visto isso, este trabalho teve o intuito de encontrar regras e padrões dentre os crimes da base de dados concedida pela Polícia Militar de Minas Gerais da cidade de Divinópolis, com objetivo de auxiliar na segurança da cidade, agilizando as futuras tomadas de decisões e estratégias.

Para encontrar essas regras e padrões, foram feitas buscas por Regras de Associação e realizadas também análises estatísticas. Em busca das Regras de Associação, foi inicialmente aplicado o algoritmo Apriori em *clusters* gerados pelo algoritmo K-means com objetivo de encontrar regras que estivessem em um subconjunto de dados mais homogêneo. Em seguida, foi aplicado o Apriori sobre a base integral para encontrar as regras que representam toda a base de dados. As análises estatísticas foram realizadas para três tipos de ocorrências específicos, sendo eles: "Roubo", "Furto" e crimes contra a mulher; conforme descrito e justificado nos capítulos anteriores.

Analisando os resultados alcançados, pode-se dizer que o trabalho alcançou os objetivos almejados. Foram encontrados conhecimentos da base de dados na forma de Regras de Associação, como as regras: "Armas de Fogo" => "Sem lesão aparente", que indica que as armas de fogo são utilizadas na maioria das vezes apenas como instrumentos de intimidação no ato do crime não gerando lesões físicas; "Noite, Roubo" => "Vítima do sexo Masculino" que talvez ocorra devido o fato de os homens se sentirem mais seguros em andar sozinhos a noite na rua, sendo assim vítimas mais frequentes nesse horário para esse tipo de crime. Além disso também foram observados certos padrões nas análises gráficas, como nos crimes contra as mulheres, onde se verificou que a maior parte das ocorrências de agressão e lesão, com vítima do sexo feminino, ocorrem durante os finais de semana, sugerindo que sejam casos de violências doméstica visto que é quando as famílias estão reunidas.

Entretanto, mesmos com os resultados encontrados, ainda assim não foram identificados conhecimentos que possam ser considerados tão inovadores, no entanto, todo o trabalho foi pautado por uma metodologia de trabalho bem definida e seguindo todas etapas de um trabalho científico. Com o intuito de trazer melhorias a este trabalho, sugere-se como trabalhos futuros, a ampliação da base de dados, trazendo dados de um período maior de tempo, já que quanto maior a base de dados histórica, maiores são as chances de conseguir descobrir novos conhecimentos e também de reforçar os padrões encontrados no trabalho atual. Ademais, sugere-se ainda buscar novas bases de dados que possam ser integradas a essa, trazendo um volume maior de dados. Finalmente, sugere-se trabalhar com dados de latitude e longitude de modo a criar um mapa com base nas ocorrências criminais presentes na base de dados.

Referências

Atlas da Violência: Divinópolis ocupa a 13ª posição entre as cidades mineiras com a maior taxa de homicídios. (2019). G1 Centro-Oeste. https://g1.globo.com/mg/centro-oeste/noticia/2019/08/06/atlas-da-violencia-divinopolis-ocupa-a-13a-posicao-entre-as-cidades-mineiras-com-a-maior-taxa-de-homicidios.ghtml.

de Amo, S. (2004). Técnicas de mineração de dados. *Jornada de Atualização em Informática*. JAI – CSBC. Salvador – BA.

Decreto-Lei 2.848, de 07 de dezembro de 1940. *Código Penal*. Rio de Janeiro, 1940. http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm

Diniz, A. M. (2005). Migração, desorganização social e violência urbana em Minas Gerais. *Raega-O Espaço Geográfico em Análise*, 9.

Dornelles, J. R. W. (2017). *O que é crime*. Brasiliense.

Engel, C. L. (2020). *A violência contra a mulher*. Instituto de Pesquisa Econômica Aplicada (IPEA). http://repositorio.ipea.gov.br/handle/11058/10313

Fayyad, U. M. et al. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAIPress, The Mit Press.

Hammound, H. J. (2013). The Value of Big Data Isn't the Data. *Harvard Business Review*. https://hbr.org/2013/05/the-value-of-big-data-isnt-the.html

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, 1(6), 90-95.

Machado, F. N. R. (2018). *Big Data O Futuro dos Dados e Aplicações.* Saraiva Educação SA.

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 (pp. 113-118).

Monitor da Violência: Mesmo com queda recorde de mortes de mulheres, Brasil tem alta no número de feminicídios em 2019. *G1*. https://g1.globo.com/monitor-da-violencia/noticia/2020/03/05/mesmo-com-queda-recorde-de-mortes-de-mulheres-brasil-tem-alta-no-numero-de-feminicidios-em-2019.ghtml

Ochi, L. S., Dias, C. R., & Soares, S. S. F. (2004). Clusterização em mineração de dados. *Instituto de Computação-Universidade Federal Fluminense-Niterói*, 1, 46.

Pereira, B. L., & Brandão, W. C. (2014). ARCA: Mining Crime Patterns Using Association Rules. In *11th International Conference Applied Computing*. Porto (pp. 159-165).

Prado, K. H. J. et al. (2020). Applied intelligent data analysis to government data related to criminal incident: A systematic review. *Journal of Applied Security Research*, 15(3), 297-331.

Prado, K. H. J., & Júnior, M. C. (2020). Data Science aplicada à análise criminal baseada nos dados abertos governamentais de Minas Gerais. *Research, Society and Development*, 9(11), e36391110044-e36391110044.

Prodanov, C. C., & De Freitas, E.C. (2013). *Metodologia do trabalho científico*: métodos e técnicas da pesquisa e do trabalho acadêmico (2a ed.), Editora Feevale.

Romão, W., Niederauer, C. A., Martins, A., Tcholakian, A., Pacheco, R. C., & Barcia, R. M. (1999). Extração de regras de associação em C&T: O algoritmo Apriori. *XIX Encontro Nacional em Engenharia de Produção*, 34, 37-39.

Scalco, P. R. (2007). *Criminalidade violenta em Minas Gerais:* Uma proposta de alocação de recursos em segurança pública. Viçosa – MG.

Sevri, M., Karacan, H., & Akcayol, M. A. (2017). Crime analysis based on association rules using apriori algorithm. *International Journal of Information and Electronics Engineering*, 7(3), 99-102.

Tayal, D. K. et al. (2015). Crime detection and criminal identification in India using data mining techniques. *AI & society*, 30(1), 117-127.