# Aplicação de Mineração de Dados para Extração de Conhecimento de Crimes de Violência Doméstica: um Estudo de Caso na Cidade de Campo Grande (MS)

*Wesley Fabricio Souza Silva*

SERVIÇO DE PÓS-GRADUAÇÃO DO FACOM-UFMS

Data de Depósito:

Assinatura:_____

# Aplicação de Mineração de Dados para Extração de Conhecimento de Crimes de Violência Doméstica: um Estudo de Caso na Cidade de Campo Grande (MS)

*Wesley Fabricio Souza Silva*

**Orientador:** *Prof. Dr. Rafael Geraldeli Rossi*
**Coorientador:** *Prof. Dr. Silvano Ferreira de Araújo*

Dissertação de mestrado apresentada à Faculdade de Computação da Universidade Federal de Mato Grosso do Sul (FACOM-UFMS) como parte dos requisitos para obtenção do título de Mestre em Computação Aplicada.

**UFMS - Campo Grande**
**Julho/2023**

*A minha mãe,*
*Elizete.*

*A minha esposa e a minhas filhas,*
*Jaqueline, Gabrielly e Beatriz.*

*A Rafael Rossi e Silvano Araújo.*

# Agradecimentos

Agradeço a Deus, por fazer parte deste momento da minha vida, e a minha família, por me apoiar em todos os instantes. A Rafael Geraldeli Rossi, meu orientador, por me direcionar durante todo o desenvolvimento deste projeto de mestrado. Ao coorientador Silvano Ferreira de Araújo pelas sugestões. Também gostaria de agradecer aos professores da FACOM pelos ensinamentos nas disciplinas. Por fim, agradeço a todos que contribuíram comigo nesse trajeto percorrido.

# Resumo

A quantidade de dados gerados e armazenados vem crescendo juntamente com o aumento do poder computacional para guardá-los. A fim de que esses dados se tornem informações úteis e possam ser utilizados por empresas e pessoas na tomada de decisões, técnicas de mineração de dados podem ser aplicadas. Por meio delas, é possível encontrar informações, associações e padrões sobre os dados analisados, os quais podem servir tanto para extrair o conhecimento presente neles quanto para fazer previsões. Com isso, os órgãos da administração pública responsáveis pela segurança da população podem se beneficiar da mineração de dados para tornar suas ações de combate e prevenção ao crime mais eficientes, como pode-se notar em diversos lugares do mundo. No estado de Mato Grosso do Sul, os dados sobre ocorrências policiais são armazenados no Sistema Integrado de Gestão Operacional (SIGO). Porém, estes ainda não têm sido explorados para auxiliar na extração de conhecimento de crimes, bem como no auxílio para a tomada de decisões mais efetivas. Posto isso, o objetivo deste trabalho é o emprego de técnicas de mineração de dados para extração de conhecimento, considerando as informações armazenadas no SIGO, mais especificamente, os dados de crimes de violência doméstica registrados na cidade de Campo Grande (MS). Além disso, por questões de explicabilidade e interpretabilidade dos resultados para os tomadores de decisão, foram utilizados algoritmos de mineração de dados do tipo simbólico: regras de associação e regras de classificação. Com isso, foi possível extrair conhecimento interessante e inovador por meio das regras de associação. Por exemplo, analisando-se o fato injúria, pôde-se perceber que ele mais frequentemente ocorre com as vítimas de escolaridade superior. Já para as regras de classificação, foi obtida uma acurácia de 84%, permitindo a extração de conhecimentos como: vítimas com idade menor ou igual a 23 anos de idade registram de 1 a 3 boletins de ocorrência contra o mesmo autor.

**Palavras-chave:** mineração de dados, extração de regras, segurança pública, violência doméstica.

# Abstract

*The amount of data generated and stored has been increasing along with the computational power for storing it. For this data to become useful information for companies and people in decision-making positions, some data mining techniques may be applied. Through them, it is possible to find information, associations, and patterns regarding the analyzed data, which may serve both to extract knowledge and to make predictions. Public administration bodies responsible for public safety can benefit from data mining to enhance their actions in fighting and preventing crime, as can be seen in many places around the world. In the state of Mato Grosso do Sul, data on police occurrences are stored in the Integrated Operational Management System (SIGO). However, they have not yet been investigated for the extraction of crime knowledge or more effective decision-making. Thus, the objective of this project is the use of data mining techniques for knowledge extraction, considering the information stored in the SIGO system, more specifically, those regarding domestic violence registered in the of Campo Grande (MS) city. Furthermore, to improve the explanation and interpretation of the results for decision makers, symbolic data mining algorithms were used: association and classification rules. With them, it was possible to extract interesting and innovative knowledge from the data. For example, analyzing the insult fact, it can be observed that the victims are more often those with higher education. As for the classification rules, an accuracy of 84% was obtained, allowing the extraction of knowledge such as: victims aged higher or equal than 23 years register 1 to 3 police reports against the same author.*

**Keywords:** *data mining, rule extraction, public security, domestic violence.*

# Sumário

Lista de Figuras . . . . . xv
Lista de Tabelas . . . . . xvii
Lista de Abreviaturas . . . . . xix

**1 Introdução** **1**
1.1 Objetivos . . . . . 2
1.2 Organização do Texto . . . . . 3

**2 Conceitos e Trabalhos Relacionados** **5**
2.1 Violência Doméstica . . . . . 5
2.1.1 Ciclo da Violência Doméstica . . . . . 6
2.1.2 PROMUSE . . . . . 7
2.1.3 Dados de Crimes de Violência Doméstica . . . . . 8
2.2 Mineração de Dados . . . . . 9
2.2.1 CRISP-DM . . . . . 10
2.2.2 Tarefas de Mineração de Dados . . . . . 12
2.3 Trabalhos Relacionados . . . . . 18
2.4 Considerações Finais . . . . . 20

**3 Metodologia** **21**
3.1 Compreensão do Negócio . . . . . 21
3.2 Entendimento dos Dados . . . . . 22
3.3 Preparação dos Dados . . . . . 23
3.4 Modelagem . . . . . 24
3.5 Avaliação . . . . . 24
3.6 Implantação . . . . . 25

**4 Resultados e Discussões** **27**
4.1 Resultado das Regras de Classificação . . . . . 27
4.2 Resultados das Regras de Associação . . . . . 29

4.2.1 Análise do Conhecimento Extraído pelas Regras de Associação . . . . . 29
4.2.2 Análise Qualitativa das Regras de Associação . . . . . 31

**5 Conclusões** **35**
5.1 Principais Resultados e Contribuições . . . . . 36
5.2 Limitações e Trabalhos Futuros . . . . . 37

**Referências Bibliográficas** **43**

# Lista de Figuras

2.1 Ciclo da violência doméstica. . . . . 7
2.2 CRISP-DM. . . . . 11
2.3 Taxonomia para caracterizar as tarefas de mineração de dados. . 12
2.4 Tarefas de mineração de dados. . . . . 13

3.1 Exemplo da base de dados disponibilizada pela SEJUSP. . . . . 22

# Lista de Tabelas

4.1 Resultado das métricas de avaliação da classificação com o algoritmo J48.PART. . . . . 27
4.2 Conjuntos de itens frequentes. . . . . 29
4.3 Regras de associação cujo êxito são vítimas que têm filhos com o agressor. . . . . 30
4.4 Regras de associação cujo êxito são vítimas com nível superior. . 30
4.5 Regras de associação cujo êxito são vítimas que têm uma profissão. 31
4.6 Regras de associação que demonstram a ameaça acompanhada de injúria. . . . . 31
4.7 Resultado da avaliação das regras de associação para o conhecimento útil. . . . . 33
4.8 Resultado da avaliação das regras de associação para o conhecimento inovador. . . . . 33

# Lista de Abreviaturas

**ACC** *Accuracy*

**AM** Aprendizado de Máquina

**BO** Boletim de Ocorrência

**CAM** Centro de Atendimento à Mulher

**CEAM** Centro Especializado de Atendimento à Mulher

**CRISP-DM** *Cross Industry Standard Process for Data Mining*

**DEAM** Delegacia Especializada de Atendimento à Mulher

**DM** *Data Mining*

**DNN** *Deep Neural Networks*

**IA** Inteligência Artificial

**kNN** *k-Nearest Neighbors*

**ML** *Machine Learning*

**PROMUSE** Programa Mulher Segura

**PMMS** Polícia Militar de Mato Grosso do Sul

**SIGO** Sistema Integrado de Gestão Operacional

**SVM** *Support Vector Machines*

CAPÍTULO

# 1

# Introdução

A quantidade de dados gerados e armazenados vem crescendo juntamente com o aumento do poder computacional de guardá-los (Katal et al., 2022). Devido ao seu grande volume, para que esses dados se tornem informações úteis que possam ser utilizadas por empresas e pessoas para auxiliar em seu dia a dia, é necessário que sejam aplicadas técnicas de mineração de dados para a extração automática de padrões (Han et al., 2011; Tan et al., 2016). Esses padrões podem conter informações úteis e/ou inovadoras, além de possibilitar a automatização de tarefas e o suporte à tomada de decisões (Vercellis, 2009; Garcia et al., 2019).

Por meio do uso de técnicas de mineração de dados, é possível encontrar informações, associações e padrões sobre os dados analisados (Mitra et al., 2002), os quais podem ser úteis tanto para extrair o conhecimento presente neles quanto para fazer previsões. Tais técnicas podem ser convenientes para diversas aplicações, tais como: filtragem de *spam*, previsão climática, classificação de proteínas, detecção de nódulos em imagens, ações de *marketing* direcionadas, identificação de possíveis crimes ou fraudes financeiras e identificação de padrões nos gastos públicos.

Os órgãos da administração pública responsáveis pela segurança da população podem se beneficiar da mineração de dados para tornar suas ações de combate e prevenção ao crime mais eficientes. Ao analisar os dados sobre ocorrências (como local, data e hora em que ocorreram), é possível encontrar padrões e agir pontualmente para reduzir o número de eventos (Silva, 2011). O Departamento de Polícia da cidade de Richmond, Estados Unidos, por exemplo, conseguiu diminuir em 47% as queixas de tiros aleatórios e aumentar em 246% o número de armas apreendidas. Isso foi possível depois de o órgão

analisar os dados coletados ao longo dos anos e, com base neles, antecipar a hora e o local dos tiros aleatórios tradicionalmente associados à véspera de Ano Novo[1]. A mineração de dados também foi utilizada pelo Departamento de Polícia da cidade de Manchester, Reino Unido, onde resultados significativos foram alcançados, tais como: redução de 24% nos crimes de roubo, diminuição de 13% nos de arrombamento e 34% nos de roubo de veículos[2].

No estado de Mato Grosso do Sul, os dados sobre ocorrências criminais são armazenados no Sistema Integrado de Gestão Operacional (SIGO)[3]. O SIGO concentra informações de todas as forças de segurança pública do estado em um único banco de dados. Vale ressaltar que não há no Mato Grosso do Sul um departamento especializado para análise dos dados, tampouco o emprego de técnicas de mineração de dados para a extração de conhecimento e suporte na tomada de decisões.

Atualmente, com o advento da Lei Maria da Penha (Lei 11.340/2006)[4], que cria mecanismos para coibir a violência doméstica e familiar contra a mulher, os números de registros dos crimes dessa natureza no SIGO tiveram um grande aumento. Em Campo Grande, por exemplo, só no ano de 2021, foram registradas 80.683 ocorrências de todos os tipos de crimes nas delegacias da capital, sendo que, desses, quase 10% equivalem somente aos de violência doméstica, uma média de 508 ocorrências por mês[5]. Todavia, devido à quantidade enorme de dados armazenados, torna-se inviável uma análise manual por busca de padrões.

## 1.1 Objetivos

O objetivo geral deste trabalho é o emprego de técnicas de mineração de dados para a extração de padrões dos crimes de violência doméstica, utilizando dados de registros na cidade de Campo Grande (MS), armazenados no SIGO, a fim de gerar conhecimento e apoiar a tomada de decisão das autoridades responsáveis.

Os objetivos específicos são listados na sequência.

- Aplicar algoritmos de mineração de dados do tipo simbólico (por exemplo, regras de associação e regras de classificação) para a extração de regras interpretáveis acerca dos padrões das vítimas que registram mais ou que registram menos boletins de ocorrência contra o mesmo agressor.

[1] https://policeandsecuritynews.com/2017/07/26/data-mining-law-enforcement/
[2] https://www.ironsidegroup.com/
[3] http://www.sigo.ms.gov.br/
[4] http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm
[5] http://estatistica.sigo.ms.gov.br/

- Apoiar o Programa Mulher Segura (PROMUSE) na classificação do grau de risco das vítimas a partir do perfil de reincidência em ocorrências.

## 1.2 Organização do Texto

Esta dissertação está dividida da seguinte forma: no Capítulo 2, é apresentado o embasamento teórico dos conceitos envolvidos neste trabalho. Em seguida, no Capítulo 3, são apresentados os detalhes do método de pesquisa utilizado para se alcançar os objetivos propostos. Já no Capítulo 4 são apresentados os resultados alcançados por meio da mineração de dados. Por fim, no Capítulo 5, as conclusões e os trabalhos futuros são apresentados.

CAPÍTULO

# 2

# Conceitos e Trabalhos Relacionados

Neste capítulo, são apresentados os conceitos-base que serão utilizados no trabalho, como a caracterização da violência doméstica, o processo de mineração de dados, e as técnicas do processo que se pretende empregar neste trabalho. Também serão apresentados estudos relacionados a este projeto.

## 2.1 Violência Doméstica

A superação da violência doméstica é um dos grandes desafios das políticas públicas no Brasil (Cerqueira et al., 2015; Avila, 2021). Sancionada em 7 de agosto de 2006, a Lei nº 11.340[1], conhecida como Lei Maria da Penha, objetiva proteger a mulher da violência doméstica e familiar (Dias, 2007). A lei recebeu esse nome devido à luta de Maria da Penha por reparação e justiça e é um importante marco na efetivação da política para as mulheres (Pougy, 2010). Trata-se de uma legislação especial cujo objetivo é criar mecanismos para coibir e prevenir a violência doméstica e familiar contra a mulher. A lei também define cinco formas de violência doméstica, conforme descrito a seguir.

1. **Violência física:** ações que ofendem a integridade física, como bater ou espancar, empurrar, atirar objetos na direção da mulher, sacudir, chutar, apertar, queimar, cortar e/ou ferir.

2. **Violência psicológica:** ações que causam danos emocionais e diminuição da autoestima ou que visam a degradar ou a controlar seus compor-

[1] http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm

tamentos, suas crenças e suas decisões, mediante ameaça, constrangimento, humilhação, manipulação, isolamento, vigilância constante, perseguição contumaz, insulto, chantagem, violação de sua intimidade, ridicularização, exploração, limitação do direito de ir e vir e/ou qualquer outro meio que cause prejuízo à saúde psicológica e à autodeterminação.

3. **Violência sexual:** ações que forçam a mulher a fazer, manter e/ou presenciar ato sexual sem que ela queira, por meio de força, ameaça e/ou constrangimento físico ou moral.

4. **Violência patrimonial:** ações que envolvem a retirada de dinheiro conquistado pela mulher com seu próprio trabalho, assim como a destruição de qualquer patrimônio, bem pessoal e/ou instrumento profissional.

5. **Violência moral:** ações que desonram a mulher diante da sociedade com mentiras e/ou ofensas. São exemplos: acusá-la publicamente de ter praticado crime, xingá-la diante dos amigos, incriminá-la de algo que ela não fez e propagar informações inverídicas sobre ela para os outros.

Normalmente, a violência não acontece da noite para o dia. Ela vai se desenvolvendo aos poucos, de forma sutil e sorrateira. Na próxima seção, será apresentada a caracterização do ciclo da violência doméstica.

### *2.1.1 Ciclo da Violência Doméstica*

A violência contra a mulher passa por um ciclo com várias fases, que se repetem e podem durar muitos anos, terminando, por vezes, no feminicídio (Veiga, 2021). Na Figura 2.1, é apresentada uma ilustração do ciclo da violência doméstica. Tal ciclo foi identificado em 1979 pela psicóloga norte-americana Lenore Walker e possui 3 fases:

1. **Fase da tensão.** É quando, entre o casal, ocorre bate-bocas, atritos, insultos e hostilidades devido a desigualdades de gênero.

2. **Fase da violência.** O agressor golpeia a mulher com mãos e/ou pés ou usando objetos de corte e/ou contundentes.

3. **Fase da lua de mel.** O agressor apresenta desculpas para justificar-se, pede perdão, dá presentes e faz promessas de que vai mudar. A mulher acredita por querer manter a relação, em razão de vários fatores, como os sociais e as dependências psicológica, emocional e econômica. Por um tempo, a situação se acalma, mas, como os conflitos de gênero não foram resolvidos, tudo recomeça: a fase da tensão e então a fase da violência, que vai se tornando cada vez mais intensa, enquanto o período entre uma e outra se encurta. Muitas vezes, o fim desse ciclo é a morte da vítima.

Figura 2.1: Ciclo da violência doméstica.


Fonte: Albani (2021)

Levando-se em conta fatores como a necessidade do rompimento do ciclo da violência doméstica, o alto número de registros envolvendo agressões, a dificuldade de combatê-las por meio do policiamento ostensivo, e o fato de frequentemente haver inibição por parte da mulher que sofre violência doméstica, foi criado no estado de Mato Grosso do Sul o PROMUSE, o qual será descrito na próxima seção.

### 2.1.2 PROMUSE

O Programa Mulher Segura (PROMUSE)[2] é um projeto instituído por meio da Portaria PMMS nº 032/2018. Tem como propósitos monitorar e proteger mulheres em situação de violência doméstica e familiar. Policiais militares devidamente capacitados realizam policiamento orientado com o objetivo de promover o enfrentamento à violência doméstica contra mulheres, por meio de ações de prevenção, visitas técnicas, conversas com vítimas, familiares e até mesmo com os agressores, fazendo os encaminhamentos pertinentes aos órgãos da rede municipal de atendimento à mulher em situação de violência. Esse tipo de policiamento é de suma importância para o enfrentamento ao crime de violência doméstica, pois ela ocorre majoritariamente nas residências

[2] https://www.naosecale.ms.gov.br/promuse/

das vítimas, o que dificulta o combate por meio do policiamento ostensivo.

O PROMUSE possui três eixos orientadores:

- Ações e campanhas no âmbito da prevenção primária, em especial, ações educativas voltadas para a prevenção à violência doméstica e familiar;

- Ações de prevenção secundária, com foco nas famílias em contexto de violência doméstica e familiar, por meio de policiamento ostensivo, fiscalizações das medidas protetivas e visitas solidárias;

- Articulação com os órgãos que compõem a rede de enfrentamento à violência contra a mulher, bem como com entidades não governamentais e sociedade civil.

Atualmente, o PROMUSE não conta com um modelo baseado em dados para a classificação do grau de risco das vítimas. Tal grau é analisado por meio de um questionário de risco elaborado pela promotoria e respondido manualmente pela vítima após contato com os policiais, em visita técnica da equipe. Ao final das respostas, é classificado em **Risco Moderado** ou **Risco Considerado**. Questionários categorizados como Risco Considerado são novamente encaminhados à vara de violência doméstica.

### *2.1.3 Dados de Crimes de Violência Doméstica*

Os dados usados nesta pesquisa foram extraídos do SIGO, que é disponibilizado por uma empresa terceirizada e utilizado pela Secretaria de Estado de Justiça e Segurança Pública (SEJUSP) para armazenar dados das ocorrências atendidas pela Polícia Civil, pela Polícia Militar, pelo Corpo de Bombeiros e por outras unidades de segurança do estado. Tendo como objetivo dinamizar o atendimento à população, o SIGO substituiu o antigo modelo de registro de ocorrência manual. Vale ressaltar que a execução deste projeto de mestrado não violará nenhuma restrição imposta pela Lei nº 13.709/2018[3] — Lei Geral de Proteção de Dados (LGPD). A ferramenta que possibilita a extração dos dados no SIGO já satisfaz às normas da LGPD, e eles já são todos anonimizados no momento da extração no sistema. Segundo Maldonado (2019), a LGPD não considera um dado anonimizado — ou seja, aquele relativo ao titular que não possa ser identificado — como um dado pessoal ou que resulta na inaplicabilidade da legislação em estudo para tal tipo de dado.

[3] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

## 2.2 Mineração de Dados

A mineração de dados (*data mining*) pode ser entendida como um processo de descoberta de conhecimento útil em grandes volumes de dados, cujo principal objetivo é identificar padrões presentes neles (Hand et al., 2001). A partir dos padrões descobertos, têm-se condições de gerar conhecimento útil para um processo de tomada de decisão ou, ainda, de automatizar processos envolvendo dados (Mariano et al., 2021).

A mineração de dados trata da aplicação de técnicas, implementadas por meio de algoritmos computacionais, capazes de receber, como entrada, um conjunto de fatos ocorridos no mundo real e devolver, como saída, um padrão de comportamento, o qual pode ser expresso, por exemplo, como uma regra de associação, uma função de mapeamento ou uma modelagem de perfil (Silva et al., 2017). Vale ressaltar que buscar padrões em um grande volume de dados é uma tarefa complexa e dispendiosa, devido à capacidade armazenada e ao número de variáveis que deve ser levado em consideração.

A mineração de dados, quando aplicada à segurança pública, tem se mostrado bastante eficiente. De acordo com Brayne (2017) e Ferguson (2017), os departamentos de polícia dos EUA nas últimas décadas têm usado cada vez mais *softwares* preditivos para atingir potenciais vítimas e infratores e prever quando e onde crimes futuros provavelmente ocorrerão. Até a década de 1970, o policiamento nesse país era amplamente reativo, envolvendo patrulhas aleatórias, respostas a chamadas para o 911 e investigações. No entanto, profissionais e pesquisadores observaram que essas estratégias tiveram pouco efeito sobre as taxas de criminalidade. Com o tempo, o policiamento mudou para práticas proativas e orientadas por dados (Braga & Weisburd, 2010; Sherman et al., 1989).

Ainda sobre a mineração de dados na área de segurança pública, ela pode ser utilizada também para: determinar os locais com índice mais alto de criminalidade, definir perfis de vítimas e criminosos, identificar a existência de quadrilhas e *serial killers*, detectar em quais dias da semana ocorrem mais delitos e até mesmo as suas causas, entre outras possibilidades. Como mencionado, um dos benefícios que a mineração de dados aplicada à segurança pública pode trazer é a possibilidade de definir o perfil das vítimas, função esta que será utilizada neste trabalho a fim de subsidiar o PROMUSE no atendimento prioritário das vítimas em estado mais vulnerável.

A mineração de dados é composta por uma sequência de etapas, de forma a alcançar o sucesso da sua aplicação. Há diferentes métodos na literatura que definem os passos para a execução de um processo de mineração de dados. Neste trabalho, foi adotado o método *Cross Industry Standard Process for Data*

*Mining* (CRISP-DM) (Lopez, 2021), o qual será detalhado na próxima seção.

### 2.2.1 CRISP-DM

Proposto em 1996 por um consórcio de empresas, o CRISP-DM é dividido em 6 fases, cuja finalidade é definir os passos a se seguir em um projeto de mineração de dados. São elas: compreensão do negócio, entendimento dos dados, pré-processamento dos dados, modelagem, avaliação e implantação, conforme detalhadas a seguir (Azevedo & Santos, 2008):

1. **Compreensão do negócio.** Esta é a primeira fase do ciclo, quando são definidos os objetivos do projeto e as necessidades da empresa ou do projeto em análise. Por isso, é necessário que todos os envolvidos estejam bem-informados e completamente alinhados. Deve-se formalizar os objetivos específicos a serem alcançados e os critérios para averiguar se obtiveram êxito ou não. A partir disso, propõe-se uma tarefa de mineração de dados que possa contribuir para a consecução dessas metas. Os recursos, tanto os tecnológicos, como dados, infraestrutura e *software*, quanto os humanos necessários para a realização da tarefa de mineração devem ser mapeados e avaliados quanto à sua disponibilidade. Nessa etapa, é elaborado o plano do projeto, com a especificação dos passos a serem executados e a definição do problema.

2. **Entendimento dos dados.** Nesta fase, realiza-se a seleção dos dados disponíveis e a sua análise, buscando maior familiaridade com eles. A análise consiste em identificar problemas de qualidade nos dados; identificar os tipos de atributos; descrever os dados em termos de formato, quantidade de registros e atributos; estimar a distribuição dos atributos; e verificar a existência de relacionamentos entre pares de atributos.

3. **Pré-processamento dos dados.** Aqui, os dados são tratados, visando a torná-los adequados à aplicação dos algoritmos que serão usados para a indução de modelos. As principais tarefas a serem executadas nesta fase são a seleção dos atributos relevantes e a limpeza dos dados, seguida por sua construção, integração e formatação, destinando-se à entrada nos algoritmos de indução a serem utilizados. São também tomadas decisões relativas à aplicação de técnicas para remoção de ruído ou de dados espúrios, estratégias para lidar com valores faltantes, criação de atributos derivados e de novos registros, integração de tabelas, se existirem, e discretização dos dados numéricos, se necessário.

4. **Modelagem.** A modelagem consiste na aplicação de técnicas que objetivam encontrar padrões e descobrir conhecimento. Faz-se uso de algorit-

mos de aprendizado de máquina, que se destinam a melhorar o desempenho em tarefa ou extrair padrões com base em exemplos. Os exemplos, nesse caso, são fornecidos pela etapa anterior.

5. **Avaliação.** A avaliação consiste em testar a efetividade do modelo aplicado. Normalmente, essa análise é baseada em indicadores e métricas comparativas. Trata-se de uma validação da adequação dos tratamentos aplicados aos dados e da modelagem escolhida.

6. **Implantação.** Na implantação, os resultados alcançados são colocados à disposição do usuário, com a finalidade de melhorar os processos de negócio.

Na Figura 2.2, são apresentadas as etapas definidas pelo modelo CRISP-DM, bem como a interação entre elas. As setas indicam as dependências mais importantes e as interações entre as etapas. A sequência das fases não é rigorosa. Na verdade, a maioria dos projetos se move para frente e para trás entre as fases conforme necessário.

Figura 2.2: CRISP-DM.


Fonte: Lemos (2020)

O processo de mineração de dados ou os padrões extraídos podem ser aplicados para diferentes propósitos, os quais são divididos como tarefas de mineração de dados, detalhadas a seguir.

### 2.2.2 Tarefas de Mineração de Dados

Na literatura, são encontradas diferentes taxonomias para caracterizar as tarefas de mineração de dados. Para Fayyad et al. (1996) e Tan et al. (2019), a taxonomia é dividida em dois níveis. No primeiro, as tarefas de mineração de dados são subdivididas em **preditivas** e **descritivas**. As tarefas preditivas envolvem o uso dos atributos de um conjunto de dados para a realização da previsão do valor futuro da variável, também sendo útil para a tomada de decisão (Rezende, 2003). Já as tarefas descritivas têm o objetivo de encontrar padrões que descrevem os dados de maneira que o ser humano possa interpretá-los, não sendo objetivo das tarefas descritivas a previsão de valores (Silva et al., 2017).

Já no segundo nível, as tarefas preditivas e descritivas são especializadas. No conjunto das preditivas, os autores inserem classificação e regressão. No conjunto das descritivas, as especializações se dividem em agrupamento, associação, sumarização, modelagem de dependências e detecção de anomalias (Mariano et al., 2021; Silva et al., 2017; Tan et al., 2019). Na Figura 2.3 é apresentada uma ilustração da taxonomia de tarefas de mineração de dados segundo a divisão de Fayyad et al. (1996) e Tan et al. (2019).

Figura 2.3: Taxonomia para caracterizar as tarefas de mineração de dados.

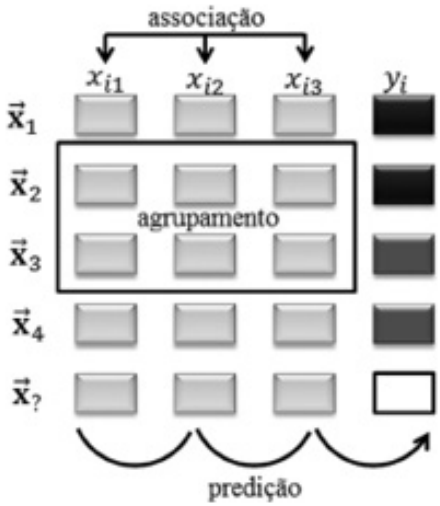

Fonte: Baseado em Tan et al. (2019)

Han et al. (2011) também se baseiam no primeiro nível da taxonomia já apresentada; porém, o segundo nível da taxonomia seguida por esses autores difere levemente do da taxonomia de Fayyad et al. (1996). Para Han et al. (2011), as tarefas de mineração de dados no segundo nível se dividem em: classificação e regressão; mineração de padrões frequentes, associações e correlações (que correspondem a um dos objetivos da "modelagem de dependências" na primeira taxonomia); análise de grupos (equivalente a "agru-

pamento"); e análise de *outliers* (similar à detecção de desvios). Ainda, no segundo nível da taxonomia, para o caso de tarefas descritivas, Maimon (2014) enumera duas tarefas adicionais: resumo linguístico e visualização. Interessante notar que esses autores, na realidade, apresentam uma taxonomia em três níveis para "paradigmas de mineração de dados", em que um nível ainda mais alto é definido, dividido em paradigmas de verificação e de descoberta. No primeiro, tarefas de teste de hipótese, análise de variância e teste de distribuição são incluídas.

Silva et al. (2017) trazem uma visão mais abrangente sobre as tarefas de mineração de dados, enumerando as três grandes tarefas: predição, agrupamento de dados e associação (ou descoberta de regras de associação). Elas podem assumir mais de uma variação e dar origem a subtarefas. Por exemplo, a detecção de anomalias pode ser resolvida a partir de uma análise de agrupamento de dados. Na Figura 2.4, é apresentada uma ilustração da visão de tarefas de mineração de dados de Silva et al. (2017).

Figura 2.4: Tarefas de mineração de dados.




Fonte: Silva et al. (2017)

Dada a gama de possibilidades de tarefas de mineração de dados que podem ser aplicadas em um conjunto de dados, neste trabalho foram utilizadas as técnicas de classificação e análise de associações, uma vez que se julgaram as mais adequadas para a extração de conhecimento, tendo em vista os objetivos definidos para este projeto. Nas próximas seções, serão detalhadas as atividades utilizadas, bem como seu esquema de avaliação.

#### 2.2.2.1 Classificação

Conforme Rezende (2003), a tarefa de classificação pode ser caracterizada por uma função de aprendizado que mapeia dados de entrada, em um número finito de categorias. A classificação pode ser:

- Binária (cada exemplo pertencendo a uma dentre duas classes possíveis) (Castro & Ferrari, 2016; Tan et al., 2019);
- Multiclasse (um exemplo pode pertencer a uma dentre três ou mais classes possíveis) (Aly, 2005; Pang et al., 2008);
- Multirrótulo (um exemplo pode pertencer a uma ou mais classes dentre múltiplas classes possíveis) (Tsoumakas et al., 2010; Zhang & Zhou, 2014; Gonçalves et al., 2018).

A classificação poderá ainda apresentar algoritmos baseados em diferentes paradigmas (Russell & Norvig, 2004; Rossi, 2011; Corcovia & Alves, 2019), tais como os apresentados a seguir.

1. **Simbólico:** os classificadores baseados no paradigma simbólico produzem modelos facilmente interpretáveis. Buscam aprender construindo representações simbólicas de um conceito por meio da análise de exemplos e contraexemplos desse conceito. As representações simbólicas estão tipicamente na forma de alguma expressão lógica, árvore de decisão, regra de produção ou rede semântica.
2. **Baseado em instâncias:** a classificação baseada em instâncias realiza novas classificações com base em casos similares cuja classe é conhecida, assumindo que o novo caso terá a mesma classe. Um dos classificadores baseados em instâncias mais utilizado é o *k Nearest Neighbors* ($k$-NN).
3. **Conexionista:** a classificação conexionista é baseada na simulação dos componentes do cérebro, e seu principal exemplo é o das redes neurais.
4. **Estatístico:** no paradigma estatístico, utiliza-se um modelo estatístico que encontre uma hipótese com boa aproximação do conceito a ser induzido. O aprendizado consiste em encontrar os melhores parâmetros para os modelos, que podem ser paramétricos (quando fazem alguma suposição sobre a distribuição dos dados) ou não paramétricos (quando não fazem suposição sobre a distribuição dos dados). Dentre os modelos estatísticos utilizados em aprendizagem de máquina, podemos destacar os modelos Bayesianos.

Os algoritmos de aprendizado baseados em regras pertencem ao paradigma de aprendizado simbólico, produzindo modelos facilmente interpretáveis (Tan et al., 2019). A classificação realizada por eles faz uso de um conjunto de regras do tipo SE-ENTÃO, em que os classificadores geram um conjunto de regras de forma que todos os exemplos de um conjunto de treinamento sejam classificados por ao menos uma das regras (Rossi, 2015).

O lado esquerdo da regra, que contém um conjunto de condições, é chamado de antecedente da regra, e o lado direito, que contém a classe, é chamado de consequente da regra. Em uma classificação automática de textos baseada em regras de classificação, por exemplo, quando um documento satisfaz todas as condições de uma regra, é dito que o documento "dispara" a regra. A cobertura de uma regra refere-se ao conjunto de documentos disparados por ela. Para cada classe, são geradas regras que cobrem corretamente todos os exemplos disparados por elas (Rossi, 2015).

Para estimar a performance de um classificador em um cenário real, pode-se fazer uso de esquemas de avaliação, como o *holdout* e o *k-fold cross-validation* (Cunha, 2019). No método *holdout*, a base de dados original é dividida em duas partes, sendo uma para treinamento e outra para teste. Um modelo de classificação é então induzido a partir do conjunto de treinamento, e seu desempenho é avaliado no conjunto de teste (Tan et al., 2019). Geralmente, a proporção dessa divisão é de 2/3 para treino e de 1/3 para teste. No *k-fold cross-validation*, a amostra de dados $D$ é dividida em $k$ partes de tamanhos semelhantes, e, a cada iteração, os dados da amostra $D_k$ são utilizados para o teste e o restante para treinamento, de modo que, ao final do processo, todos os dados terão sido utilizados como teste. A escolha do valor de $k$ é uma importante etapa para esse processo: um valor de $k$ baixo pode diminuir em muito o número de exemplos de treinamento, enquanto um valor de $k$ alto pode ocasionar um alto gasto computacional e, consequentemente, demandar maior tempo para processamento, sendo mais comum o uso de $k = 10$ (Cunha, 2019; Tan et al., 2019).

Quanto às medidas de avaliação do modelo de classificação, o mesmo pode ser avaliado por meio das métricas a seguir (Rossi, 2011).

- **Precisão** — avaliação dos acertos para a classe positiva em relação a todos os exemplos que o modelo classificou como pertencentes à classe. Seguindo a fórmula:

$$precisao = \frac{VP}{VP + FP}, \tag{2.1}$$

  em que $VP$ são os elementos que o modelo classificou como positivos e que são verdadeiramente positivos e $FP$ são elementos que o modelo

classificou como positivos, mas que foram rotulados como negativos.

- **Revocação** — avaliação dos acertos para uma determinada classe em relação a todos os exemplos que verdadeiramente pertenciam à classe. Seguindo a fórmula:

$$revocacao = \frac{VP}{VP+FN}, \tag{2.2}$$

em que *FN* são elementos que o modelo classificou como negativos, mas que foram rotulados como positivos.

- **Acurácia** — a acurácia (*accuracy* ou ACC) é considerada uma das métricas mais simples e importantes. Ela avalia unicamente o percentual de acertos, ou seja, pode ser obtida pela razão entre a quantidade de acertos e o total de entradas:

$$acuracia = \frac{Total\ de\ acertos}{Total\ de\ itens} \tag{2.3}$$

Utilizando como base a matriz de confusão, podemos obter a acurácia pela fórmula (Ferrari & Silva, 2017):

$$acuracia = \frac{VP+VN}{VP+FN+VN+FP} \tag{2.4}$$

- **F1-score** — F–*measure*, F-*score* ou *score* $F_1$ é uma média harmônica calculada com base na precisão e na revocação. Ela pode ser obtida com base na equação (Ferrari & Silva, 2017):

$$F_1 = 2 * \frac{precisao * revocacao}{precisao + revocacao} \tag{2.5}$$

#### 2.2.2.2 Regras de associação

De acordo com Silva et al. (2017), a descoberta de regras de associação é o processo de analisar os relacionamentos existentes entre atributos de uma base de dados transacional, com o objetivo de encontrar associações ou correlações entre esses atributos. Ou seja, o quanto a presença de um conjunto de itens nos registros de uma base de dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (Agrawal & Srikant, 1994). Desse modo, o objetivo das regras de associação é o de encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados (Rossi, 2011).

As regras de associação são formadas por dois conjuntos de elementos, um **antecedente** e um **consequente**, representados na forma Antecedente

$(A) \Rightarrow$ Consequente $(B)$, e interpretada da seguinte maneira: A ocorrência de $A$ implica na ocorrência de $B$, ou ainda, se $A$ ocorreu, também é provável que $B$ ocorra. Em geral, a mineração de regras de associação pode ser dividida em duas etapas (Goldschmidt & Passos, 2005):

1. Encontrar todos os *itemsets* frequentes (que satisfazem à condição de suporte mínimo);
2. A partir dos *itemsets* frequentes, gerar as regras de associação (que satisfazem à condição de confiança mínima).

O suporte é simplesmente o número de transações que incluem todos os itens na parte antecedente e consequente da regra. Assim, para uma determinada regra de associação $A \Rightarrow B$, o suporte da regra mede o número total de registos de transação que contêm os conjuntos de itens $A$ e $B$. Neste caso, o suporte da regra $A \Rightarrow B$, onde $A$ e $B$ são conjuntos de itens, seria dado pela seguinte expressão:

$$sup(A \Rightarrow B) = \frac{Frequencia\ de\ A\ e\ B}{Total\ de\ T}, \tag{2.6}$$

em que o numerador diz respeito ao número de transações em que $A$ e $B$ ocorrem simultaneamente e o denominador refere-se ao número total de transações da base de dados.

A outra medida utilizada na descoberta de regras de associação é a confiança do conjunto de itens frequentes. A confiança de uma regra é dada pela seguinte fórmula:

$$conf\ (B \Rightarrow A) = \frac{sup(B \cup A)}{sup(\ B\ )}, \tag{2.7}$$

em que o numerador refere-se ao número de transações em que $B$ e $A$ ocorrem simultaneamente. O denominador refere-se à quantidade de transações em que o item $B$ ocorre. Em termos gerais a confiança mede a probabilidade condicional de ocorrer $B$ dado que ocorreu $A$. Outras medidas estatísticas também são usadas na descoberta de regras de associação, tais como (Mariano et al., 2021):

- **Convicção** — razão da frequência esperada que $B$ ocorre sem $A$, ou seja, a frequência em que a regra faz uma predição incorreta. A convicção é calculada por:

$$conv\ (B \Rightarrow A) = \frac{1 - sup(\ A\ )}{1 - conf\ (B \Rightarrow A)} \tag{2.8}$$

- ***Lift*** **(Elevação)** — usada para avaliar a dependência entre os conjuntos de itens, ou seja, quanto mais frequente torna-se $B$ quando $A$ ocorre:

$$lift\ (B \Rightarrow A) = \frac{conf(B \Rightarrow A)}{sup\ (A)} \tag{2.9}$$

A obtenção de *itemsets* frequentes para gerar regras de associações pode ser realizada utilizando diversos algoritmos (Rossi, 2011). Um dos mais conhecidos e usados para gerar *itemsets* frequentes é o *Apriori* (Imielinski et al., 1993).

Para a avaliação das regras de associação, as mesmas serão ranqueadas de acordo com as medidas objetivas já apresentadas, como suporte, confiança, elevação e convicção, e serão apresentadas a especialistas de domínio. Estes responderão a um questionário que irá avaliar o conhecimento extraído por meio de regras.

No presente estudo, as regras de associação serão utilizadas com o propósito de se extrair o padrão das vítimas que registram mais ou que registram menos boletins de ocorrência contra o mesmo agressor.

## 2.3 Trabalhos Relacionados

Nesta seção, são apresentados os trabalhos encontrados na literatura que tratam do uso de mineração de dados em bases de dados da segurança pública.

Em Silva (2011), os autores, utilizando o algoritmo *Apriori* para a tarefa de regra de associação e o suporte mínimo = 10% e a confiança mínima de 75% como medida objetiva, conseguiram extrair conhecimento da base de dados de crimes, respondendo a perguntas como: qual faixa etária entre as vítimas é mais suscetível aos crimes de homicídio (19-28 anos); quais horários há maior número de homicídios dolosos (noite, 33%, e madrugada, 32%); e o estado civil das vítimas (solteiras).

Em Wang et al. (2013), os autores, utilizando o algoritmo de detecção de padrões *Series Finder*, encontraram padrões de crimes cometidos pelo mesmo criminoso. Isso foi possível a partir da análise do chamado *modus operandi* do criminoso. O *modus operandi* é um conjunto de hábitos que o infrator segue, determinado tipo de comportamento que acaba por caracterizar um padrão. Por exemplo, o *modus operandi* para os arrombamentos inclui fatores como meio de entrada (porta da frente, porta dos fundos, janela), dia da semana, características do imóvel (apartamento, casa unifamiliar) e proximidade geográfica com outros arrombamentos. Usando um banco de dados de crimes anteriores, o *Series Finder* processa informações de maneira semelhante à de como os analistas de crime processam informações instintivamente: o *Series Finder* procura semelhanças entre os crimes cometidos e o *modus operandi* do agressor ou grupo específico que comete esses crimes. Posteriormente, rea-

liza uma classificação dos possíveis suspeitos. À medida que mais crimes são adicionados ao conjunto, o *modus operandi* torna-se mais bem definido.

Já em Sathyadevan et al. (2014), os autores, fazendo uso do algoritmo *Naive Bayes*, mediram a acurácia da classificação e a previsão de regiões com alta probabilidade de ocorrência de crimes com base em diferentes conjuntos de testes. A acurácia obtida foi de 90%. Os dados criminais vieram de várias fontes, como *sites* de notícias, *blogs*, mídias sociais e *feeds Rich Site Summary* (RSS). Foi possível observar nesse trabalho a possibilidade de uma análise não somente em bases de dados estruturadas dos órgãos de segurança pública para predição de crimes.

No estudo de Hassani et al. (2016), os autores fizeram uma revisão dos principais aplicativos de mineração de dados sendo adotados no combate ao crime. Foi observado que as técnicas de classificação são a forma mais popular de mineração de dados criminais. Observou-se ainda que o SVM, as redes neurais e a mineração de regras de associação raramente são usadas para mineração de estatísticas oficiais, enquanto na mineração de dados criminais esses métodos são extremamente populares e bem explorados.

Ivan et al. (2017) desenvolveram em seu trabalho um modelo de previsão de crimes usando o algoritmo de árvore de decisão J48, baseado no algoritmo C4.5 (Quinlan, 2014), que, quando empregado no contexto de aplicação da lei e de análise de inteligência, promete abrandar as taxas de criminalidade e é considerado o algoritmo de ML mais eficiente para a previsão de dados de crimes na literatura relacionada. Os resultados experimentais do algoritmo J48 previram a categoria desconhecida de dados de crime com uma acurácia de 94,25%.

Bharati (2018) analisaram em seu trabalho um conjunto de dados composto por vários crimes e previram o tipo de crime que pode ocorrer em uma situação futura. O conjunto de dados do crime consiste em informações como a descrição do local do crime, tipo de crime, data, hora e coordenadas precisas do local. Diferentes combinações de modelos, como classificação $k$-NN, regressão logística, árvore de decisão, floresta aleatória e métodos bayesianos, foram testadas. A classificação $k$-NN mostrou-se a melhor, com uma acurácia de aproximadamente 78%. Os pesquisadores também usaram diferentes gráficos, que ajudaram a entender as várias características do conjunto de dados de crimes de Chicago.

No trabalho de Hossain et al. (2020), propôs-se um sistema que prevê o crime analisando-se um conjunto de dados contendo registros de crimes e seus padrões. Tal sistema funciona principalmente com dois algoritmos de ML: uma árvore de decisão e $k$-NN. Técnicas como algoritmo de floresta aleatória e *Adaptive Boosting* foram utilizadas para aumentar a acurácia do modelo

de previsão. O sistema proposto foi alimentado com dados de atividades criminosas por um período de 12 anos em São Francisco, Estados Unidos. Usando métodos de subamostragem e sobreamostragem juntamente com o algoritmo de floresta aleatória, a acurácia foi aumentada para 99,16%.

Safat et al. (2021) analisaram oito algoritmos de aprendizado de máquina para obter previsões precisas nos conjuntos de dados sobre crimes de Chicago e Los Angeles. Os algoritmos implementados no estudo foram: regressão logística, árvore de decisão, floresta aleatória, MLP, *Naive Bayes*, SVM, *XGBoost* e *k*-NN. O *XGBoost* obteve um desempenho melhor do que os outros algoritmos, com 94% e 88% de acurácia nos conjuntos de dados de Chicago e Los Angeles, respectivamente. O conjunto de dados da cidade de Chicago continha o histórico de crimes (relatos e fatores sociais) de 2001 a novembro de 2019; já o conjunto de dados da cidade de Los Angeles continha o histórico criminal de 2010 a 2018.

No trabalho de Aziz et al. (2022), os autores usaram diferentes algoritmos de regressão, a saber, Regressão Linear (LR), Regressão de Árvore de Decisão (DTR), *Support Vector Regression* (SVR) e *Random Forest Regression* (RFR) para construir modelos de regressão. O modelo tenta prever a quantidade de crimes que ocorrerá em diferentes regiões da Índia. A base de dados continha informações criminais de 2001 a 2012. O $R^2$ ajustado e o Erro Percentual Médio Absoluto foram utilizados para avaliar e comparar os modelos de regressão propostos. O *Random Forest Regression* (RFR) se encaixou melhor ao modelo de previsão. Ele obteve um valor $R^2$ ajustado de 96% e um erro de 20%. Para a previsão de contagem de crimes de roubos, o modelo baseado no RFR alcançou um valor de $R^2$ de 96% e um erro de 16%.

## 2.4 Considerações Finais

Neste capítulo, foram apresentadas definições acerca do crime de violência doméstica e seu ciclo, detalhes sobre o PROMUSE e a fonte de dados que será utilizada para a pesquisa. Foi apresentado também o processo de mineração de dados, bem como técnicas do processo que se pretende empregar neste trabalho. Por fim, foram apresentados trabalhos relacionados e as principais técnicas utilizadas nesses trabalhos.

No próximo capítulo, serão apresentados os detalhes do método de pesquisa adotado para alcançar os objetivos desta dissertação.

CAPÍTULO

# 3

# Metodologia

Este capítulo contém os passos seguidos para o desenvolvimento do trabalho e para o alcance dos objetivos propostos. A metodologia CRISP-DM foi utilizada para guiar o processo de extração de conhecimento da base de dados.

## 3.1 Compreensão do Negócio

Nos formulários dos boletins de ocorrência cadastrados no banco de dados do SIGO, registram-se todos os dados relacionados ao crime (fato, localidade, data, horário e outros), ao criminoso e à vítima (sexo, idade, características físicas, estado civil, entre outros). Para a pesquisa, foi adquirida, por meio da Secretaria de Justiça e Segurança Pública de MS, uma base de dados contendo registros de crimes de violência doméstica na cidade de Campo Grande (MS), compreendendo os meses de março, abril e maio do ano de 2021.

O objetivo do negócio é extrair o padrão das vítimas que registram mais ou que registram menos boletins de ocorrência contra o mesmo agressor e, com isso, tentar entender o que as leva a permanecer no ciclo da violência doméstica. O conhecimento novo obtido deve ser capaz de subsidiar o PROMUSE na classificação das vítimas que registram até três boletins de ocorrência ou vítimas que registram mais de três boletins de ocorrência, de forma a priorizar o atendimento.

Para alcançar o objetivo do negócio, serão utilizados algoritmos de mineração de dados do tipo simbólico: regras de associação e regras de classificação. Esses deverão ser capazes de extrair regras interpretáveis acerca dos padrões das vítimas que registram mais ou que registram menos boletins de ocorrência. Com base na complexidade do problema, a performance de classificação

deverá ser de no mínimo 70%. Já as regras de associação serão ranqueadas de acordo com medidas objetivas e serão apresentadas a especialistas de domínio. Estes responderão a um questionário, a fim de avaliar a validade e a inovação do conhecimento extraído por meio das regras de associação.

## 3.2 Entendimento dos Dados

A base de dados foi disponibilizada no formato *Comma Separated Values* (CSV) e possui 1361 registros. Para o estudo, foram descartadas as ocorrências em que o autor não fosse o cônjuge, visto que pai, irmão, tio e outros pertencentes ao seio familiar poderiam caracterizar-se também como autores; todavia, atributos como FILHOS JUNTOS, TEMPO JUNTOS ficariam sem sentido, restando 555 registros para a análise, conforme a Figura 3.1.

Figura 3.1: Exemplo da base de dados disponibilizada pela SEJUSP.

| NR BO | FATO | FILHOS JUNTOS | TEMPO JUNTOS | DATA DO FATO | HORA DO FATO | IDADE NA DATA DO FATO | TRABALHA | ESCOLARIDADE | TIPO DE LOCAL DO FATO | QTD REG ANT |
|---|---|---|---|---|---|---|---|---|---|---|
| 2500/2021 1DEAM | INJURIA (VIOLENCIA DOMESTICA) | S | 204.00 | sexta-feira | TARDE | 37 | S | SUPERIOR COMPLETO | RESIDENCIA | 1.00 |
| 2499/2021 1DEAM | LESAO CORPORAL DOLOSA (VIOLENCIA DOMESTICA), A... | S | 7.00 | segunda-feira | MANHA | 26 | N | SUPERIOR INCOMPLETO | RESIDENCIA | 1.00 |
| 2498/2021 1DEAM | AMEACA (VIOLENCIA DOMESTICA) | N | 48.00 | segunda-feira | TARDE | 26 | S | FUNDAMENTAL INCOMPLETO | VIA URBANA | 5.00 |
| 2494/2021 1DEAM | INJURIA (VIOLENCIA DOMESTICA) | N | 12.00 | quinta-feira | NOITE | 40 | S | SUPERIOR COMPLETO | RESIDENCIA | 1.00 |
| 2493/2021 1DEAM | AMEACA (VIOLENCIA DOMESTICA), INJURIA (VIOLENC... | N | 48.00 | quarta-feira | NOITE | 29 | N | NaN | RESIDENCIA | 2.00 |
| 2490/2021 1DEAM | LESAO CORPORAL DOLOSA (VIOLENCIA DOMESTICA), I... | S | 42.00 | quinta-feira | MANHA | 25 | S | MEDIO COMPLETO | RESIDENCIA | 2.00 |

Fonte: Autoria própria.

Os registros contêm as informações a seguir.

- **NR BO:** número que identifica o registro da ocorrência no SIGO;
- **FATO:** atributo que descreve a tipificação penal do crime;
- **FILHOS JUNTOS:** atributo que descreve se a vítima tem ou não filhos com o agressor;
- **TEMPO JUNTOS:** atributo que descreve em meses há quanto tempo a vítima está com o agressor;
- **DATA DO FATO:** atributo que descreve o dia da semana em que ocorreu o crime;
- **HORA DO FATO:** atributo que descreve o período do dia em que ocorreu o crime;

- **IDADE NA DATA DO FATO:** atributo que descreve a idade da vítima na data do registro;

- **TRABALHA:** atributo que descreve se a vítima trabalha ou não;

- **ESCOLARIDADE:** atributo que descreve o grau de escolaridade da vítima;

- **TIPO DE LOCAL DO FATO:** atributo que descreve o local onde a vítima sofreu a violência doméstica;

- **QTD REG ANT:** atributo que descreve a quantidade de registros do crime de violência doméstica contra um agressor que a vítima já tenha denunciado anteriormente.

As informações FILHOS JUNTOS, TEMPO JUNTOS e QTD REG ANT foram coletadas manualmente no histórico das ocorrências pelo especialista de domínio, via consulta no sistema pelo NR BO e, após a leitura dos históricos, elas foram adicionadas manualmente no arquivo disponibilizado.

Com os objetivos de se obter conhecimento do conjunto de dados e identificar atributos importantes, irrelevantes e correlacionados, além de valores ausentes, exemplos redundantes e distribuição de valores dos atributos, foi realizada uma análise exploratória sobre o conjunto de dados. Nela, foi possível identificar a falta de preenchimento nos atributos TEMPO JUNTOS, ESCOLARIDADE e TRABALHA. Dessa forma, foram escolhidas técnicas mais apropriadas para as próximas etapas.

## 3.3 Preparação dos Dados

Nesta etapa, os dados categóricos ausentes nos atributos ESCOLARIDADE e TRABALHA foram tratados com a **moda**, já os dados numéricos ausentes no atributo TEMPO JUNTOS foram tratados com a **mediana**. Posteriormente, foi feita a remoção dos atributos NR BO e TIPO DE LOCAL DO FATO (predominantemente na residência da vítima) considerados irrelevantes para o processo de extração de conhecimento. O procedimento de padronização foi realizado no atributo ESCOLARIDADE para que houvesse apenas três níveis: FUNDAMENTAL, MÉDIO e SUPERIOR. O atributo QTD REG ANT foi discretizado em: **Entre_1_e_3_registros** (vítimas com 1, 2 ou 3 registros anteriores contra o agressor) e **Mais_que_3_registros** (vítimas com mais de 3 registros contra o mesmo agressor).

## 3.4 Modelagem

Na etapa de modelagem, por questões de esclarecimento dos resultados e interpretabilidade dos modelos, foram aplicados algoritmos de mineração de dados do tipo simbólico: regras de associação e regras de classificação. A biblioteca *mlxtend* [1] foi utilizada para calcular as regras de associação sobre os dados, por meio da classe *association rules*, que fornece como saída todas as métricas que quantifiquem o grau de associação entre dois conjuntos de itens, citadas no Capítulo 2, como suporte, confiança, convicção e elevação. Para o cálculo das regras de associação, foi utilizado o suporte mínimo de 0.10 e a confiança mínima de 0.20. Esses valores foram alcançados exploratoriamente, de forma a obter um conjunto de regras cujos valores não se apresentasse infrequentes e fossem capazes de gerar conhecimento novo. O algoritmo *Apriori* foi utilizado para a extração dos *itemsets* frequentes.

Já para a extração das regras de classificação, a ferramenta de *software* utilizada foi o Weka[2], um *software* livre e de código aberto utilizado para mineração de dados, desenvolvido em Java dentro das especificações da *General Public License* (GPL). A versão utilizada para o trabalho de pesquisa foi a 3.8.6. Na etapa de classificação, foram utilizados os atributos: filhosJuntos, tempoJuntos, idade, possuiProfissao, escolaridade e qtdRegAnt, sendo que a qtdRegAnt foi o atributo-alvo. Os atributos fato, diaSemana e período não foram utilizados para classificação. A porcentagem utilizada para treino foi de 70% e para teste, 30%. O algoritmo utilizado foi o J48.PART, e os parâmetros utilizados no algoritmo foram os valores-padrão da ferramenta Weka.

## 3.5 Avaliação

Para a avaliação dos resultados das regras de classificação, foi aplicado o esquema de validação *K-fold cross-validation* com *fold* = 10 e aplicadas métricas tradicionais, como acurácia, precisão, revocação, F1 e AUC-ROC. Os resultados da precisão, revocação, F1 e AUC-ROC foram baseados na **média ponderada** (*weighted avg*). A média aritmética ponderada é semelhante à média aritmética comum. A diferença, todavia, é que na média aritmética comum todos os valores concorrem com peso igual, enquanto que no cálculo da média aritmética ponderada se leva em consideração a contribuição (peso) de cada termo, uma vez que existem termos que contribuem mais que outros (Cazorla, 2003).

Já na parte referente às regras de associação, estas foram ranqueadas de

[1] `http://rasbt.github.io/mlxtend/`
[2] `https://www.cs.waikato.ac.nz/ml/weka/index.html`

acordo com medidas objetivas, como suporte, confiança, elevação e convicção e, em seguida, apresentadas para especialistas de domínio, que avaliaram a validade e a inovação do conhecimento extraído por meio das regras de associação. Os mesmos responderam perguntas como: considerando a regra $X \rightarrow Y$, o fato injúria frequentemente ocorre com as vítimas de escolaridade superior. De 0 a 10: O conhecimento apresentado foi útil? O conhecimento apresentado foi inovador?

## 3.6 Implantação

Como resultado, o conhecimento extraído por meio de regras será disponibilizado à sociedade e às autoridades competentes para fins de conhecimento em relação aos padrões de violência doméstica, bem como para o suporte à tomada de decisões no combate a esse tipo de crime pelo PROMUSE.

CAPÍTULO

# 4

# Resultados e Discussões

Neste capítulo, são apresentados os resultados das regras de classificação e das regras de associação. Também são apresentadas as análises dos resultados e das regras extraídas.

## 4.1 Resultado das Regras de Classificação

Para a avaliação dos resultados das regras de classificação, foram aplicadas métricas tradicionais, como acurácia, precisão, revocação, F1 e AUC-ROC, os resultados obtidos com o algoritmo J48.PART são apresentadas na Tabela 4.1.

Tabela 4.1: Resultado das métricas de avaliação da classificação com o algoritmo J48.PART.

| **Métrica** | **Valor da medida** |
|---|---|
| Acurácia | 84% |
| Precisão | 79% |
| Revocação | 84% |
| F1 | 80% |
| AUC-ROC | 61% |

Fonte: Autoria própria.

Na Tabela 4.1, é possível observar que a proposta de extrair regras interpretáveis acerca dos padrões das vítimas com o uso do algoritmo J48.PART obteve resultados satisfatórios. A acurácia de 84% alcançada demonstra o percentual de acertos do modelo.

Com o objetivo de extrair regras interpretáveis acerca dos padrões das vítimas que registram mais ou que registram menos boletins de ocorrência contra

o mesmo agressor, utilizando o algoritmo J48.PART, foi possível verificar que as vítimas com idade menor ou igual a 23 anos registram de 1 a 3 boletins de ocorrência contra o mesmo autor. Os números que aparecem em parênteses significam que a regra cobre o total de 91 exemplos da base de dados e erra em somente 3:

```
idade <= 23:  Entre_1_e_3_registros (91.0/3.0)
```

Também foi possível extrair que as vítimas com pouca escolaridade (FUNDAMENTAL), sem uma profissão, sem filhos juntos e relacionando-se com seus companheiros há até 96 meses (8 anos) também estão entre as que registraram de 1 a 3 boletins de ocorrência contra o mesmo agressor:

```
filhosJuntos = N AND escolaridade = FUNDAMENTAL AND
    possuiProfissao = N AND tempoJuntos <= 96:
        Entre_1_e_3_registros (52.0)
```

Verificou-se também que as vítimas com nível fundamental de escolaridade, com profissão e que não têm filhos com o agressor também estão entre as que registraram de 1 a 3 boletins de ocorrência contra o mesmo agressor:

```
filhosJuntos = N AND escolaridade = FUNDAMENTAL AND
possuiProfissao = S: Entre_1_e_3_registros (69.0/5.0)
```

Vítimas com escolaridade média, filhos e uma profissão também estão entre as que registram de 1 a 3 boletins de ocorrência contra o mesmo agressor:

```
escolaridade = MÉDIO AND possuiProfissao = S AND filhosJuntos
        = S: Entre_1_e_3_registros (26.0/3.0)
```

Já as vítimas com o nível médio de escolaridade, que têm uma profissão, estão juntas há mais de 5 meses e sem filhos juntos também estão em sua maioria entre as vítimas que registram de 1 a 3 boletins de ocorrência contra o mesmo agressor:

```
escolaridade = MÉDIO AND filhosJuntos = N AND possuiProfissao
 = S AND tempoJuntos > 5:  Entre_1_e_3_registros (26.0/3.0)
```

Vítimas que não têm uma profissão e o tempo junto é menor ou igual a 338 meses, registram entre 1 e 3 boletins de ocorrência contra o agressor:

```
possuiProfissao = N AND tempoJuntos <= 338:
    Entre_1_e_3_registros (137.0/23.0)
```

Vítimas com idade inferior a 54 anos e sem profissão registram mais que 3 boletins de ocorrência contra o mesmo autor:

```
                possuiProfissao = N AND
idade <= 53:  Mais_que_3_registros (6.0)
```

Por último, se nenhuma das regras anteriores forem disparadas, o algoritmo direciona para a classe das vítimas que registram entre 1 e 3 boletins de ocorrência:

```
:  Entre_1_e_3_registros (148.0/37.0)
```

## 4.2 Resultados das Regras de Associação

Nesta seção, são apresentados subconjuntos das regras de associação extraídas, bem como comentários sobre o conhecimento extraído e uma análise qualitativa das regras de associação.

### 4.2.1 Análise do Conhecimento Extraído pelas Regras de Associação

Na Tabela 4.2, são apresentados os 10 *itemsets* mais frequentes. Pode-se perceber que: o tipo do local em que o fato ocorre com mais constância é a residência; a quantidade de registros contra o mesmo agressor está entre 1 e 3; o ensino é fundamental; a vítima e o agressor têm filhos juntos (neste último, é apenas um pouco mais da metade das vezes); e o fato mais recorrente é o crime de ameaça.

Tabela 4.2: Conjuntos de itens frequentes.

| **Suporte** | **Conjuntos de itens** |
|---|---|
| 0.90 | (tipoLocal=RESIDÊNCIA) |
| 0.86 | (qtdRegAnt=Entre 1 e 3 registros) |
| 0.77 | (tipoLocal=RESIDÊNCIA, qtdRegAnt=Entre 1 e 3 registros) |
| 0.65 | (escolaridade=FUNDAMENTAL) |
| 0.58 | (escolaridade=FUNDAMENTAL, tipoLocal=RESIDÊNCIA) |
| 0.57 | (fato=AMEAÇA) |
| 0.56 | (filhosJuntos=S) |
| 0.56 | (escolaridade=FUNDAMENTAL, qtdRegAnt=Entre 1 e 3 registros) |
| 0.52 | (possuiProfissao=S) |
| 0.52 | (fato=AMEAÇA, tipoLocal=RESIDÊNCIA) |

Fonte: Autoria própria.

Na Tabela 4.3, são apresentadas as regras de associação ordenadas de maneira decrescente pela elevação. Nela, são apresentadas as regras de associação cujo consequente é filhosJuntos=S, isto é, têm filhos juntos. Pode-se

perceber que, em geral: o local da violência é a própria residência da vítima; protocola-se mais de 3 registros; o tipo de violência é o de injúria; e o dia mais comum de ocorrência é no sábado, sendo o período vespertino. Também é possível observar que, em 71% das vezes em que há mais de 3 registros, a vítima tem filho com o agressor, fato que ocorre em 10% da base.

Tabela 4.3: Regras de associação cujo êxito são vítimas que têm filhos com o agressor.

| **Antecedentes** | **Consequentes** | **Suporte** | **Confiança** | **Elevação** |
|---|---|---|---|---|
| (qtdRegAnt=Mais que 3 registros) | (filhosJuntos=S) | 0.10 | 0.71 | 1.27 |
| (tipoLocal=RESIDÊNCIA, periodo=TARDE, fato=INJÚRIA) | (filhosJuntos=S) | 0.10 | 0.71 | 1.26 |
| (periodo=TARDE, fato=INJÚRIA) | (filhosJuntos=S) | 0.11 | 0.69 | 1.22 |
| (tipoLocal=RESIDÊNCIA, fato=INJÚRIA, possuiProfissao=S) | (filhosJuntos=S) | 0.14 | 0.66 | 1.17 |
| (diaSemana=sábado) | (filhosJuntos=S) | 0.10 | 0.66 | 1.17 |

Fonte: Autoria própria.

Na Tabela 4.4, são apresentadas as regras de associação também ordenadas de maneira decrescente pela elevação. Nela, são apresentadas regras cujo antecedente contém a vítima com escolaridade SUPERIOR. É possível perceber que, em aproximadamente 10% das transações, em 58% das vezes, as vítimas com escolaridade superior sofrem INJÚRIA e, em 60% das vezes, a INJÚRIA ocorre na própria RESIDÊNCIA da vítima.

Tabela 4.4: Regras de associação cujo êxito são vítimas com nível superior.

| **Antecedentes** | **Consequentes** | **Suporte** | **Confiança** | **Elevação** |
|---|---|---|---|---|
| (escolaridade=SUPERIOR, tipoLocal=RESIDÊNCIA) | (fato=INJÚRIA) | 0.10 | 0.60 | 1.23 |
| (escolaridade=SUPERIOR) | (fato=INJÚRIA) | 0.11 | 0.58 | 1.19 |

Fonte: Autoria própria.

Na Tabela 4.5, são apresentadas as regras de associação também ordenadas de maneira decrescente pela elevação. Nela, é possível observar que, em aproximadamente 15% das transações, em 70% das vezes, as vítimas com escolaridade SUPERIOR também são as que têm uma profissão. Sendo ainda que, em aproximadamente 11% das transações, em 66% das vezes, as vítimas com escolaridade SUPERIOR e que possui uma profissão são também as que registram entre 1 e 3 boletins de ocorrência contra o mesmo agressor.

Tabela 4.5: Regras de associação cujo êxito são vítimas que têm uma profissão.

| **Antecedentes** | **Consequentes** | **Suporte** | **Confiança** | **Elevação** |
|---|---|---|---|---|
| (escolaridade=MÉDIO) | (possuiProfissao=S) | 0.10 | 0.72 | 1.37 |
| (escolaridade=SUPERIOR) | (possuiProfissao=S) | 0.15 | 0.70 | 1.33 |
| (escolaridade=SUPERIOR, tipoLocal=RESIDÊNCIA) | (possuiProfissao=S) | 0.13 | 0.69 | 1.31 |
| (escolaridade=SUPERIOR, qtdRegAnt=Entre 1 e 3 registros) | (possuiProfissao=S) | 0.11 | 0.66 | 1.26 |

Fonte: Autoria própria.

Na Tabela 4.6, é possível observar a presença constante do fato AMEAÇA acompanhado do fato INJÚRIA. Isso se explica ao analisarmos a frase que frequentemente é proferida pelo agressor às vítimas:

"... Vou te matar (AMEAÇA) sua *algum palavrão* (INJÚRIA)..."

Tabela 4.6: Regras de associação que demonstram a ameaça acompanhada de injúria.

| **Antecedentes** | **Consequentes** | **Suporte** | **Confiança** | **Elevação** |
|---|---|---|---|---|
| (tipoLocal=RESIDÊNCIA, possuiProfissao=S) | (fato=AMEAÇA) | 0.12 | 0.66 | 1.15 |
| (fato=INJÚRIA, filhosJuntos=N) | (fato=AMEAÇA) | 0.13 | 0.65 | 1.15 |
| (fato=INJÚRIA, filhosJuntos=N, qtdRegAnt=Entre 1 e 3 Registros) | (fato=AMEAÇA) | 0.12 | 0.65 | 1.14 |
| (fato=INJÚRIA, filhosJuntos=N, tipoLocal=RESIDÊNCIA) | (fato=AMEAÇA) | 0.12 | 0.65 | 1.14 |

Fonte: Autoria própria.

### 4.2.2 Análise Qualitativa das Regras de Associação

As regras de associação foram apresentadas aos especialistas de domínio (efetivo policial pertencente ao PROMUSE). Participaram da avaliação 4 militares (1 major, 2 sargentos e 1 cabo), ocasião em que a validade e a inovação do conhecimento extraído por meio das regras de associação foram avaliadas. O formulário com as regras foi-lhes apresentado, conforme segue:

---

**Questão 1:** Considerando a regra $X \rightarrow Y$, pode-se perceber que o tipo do local em que o fato ocorre com mais frequência é a residência; a quantidade

de registros contra o mesmo agressor está entre 1 e 3; o ensino é fundamental; a vítima e o agressor têm filhos juntos (neste último, é apenas um pouco mais da metade das vezes) e o fato mais recorrente é o crime de ameaça.

De 0 a 10:

O conhecimento apresentado foi útil?

O conhecimento apresentado foi inovador?

---

**Questão 2:** Considerando a regra $X \rightarrow Y$, das vítimas que têm filhos com o agressor, pode-se perceber que, em geral, o local da violência é a própria residência da vítima; elas registram mais de 3 boletins de ocorrência; o tipo de violência predominante é o de injúria e o dia mais comum de ocorrências, no sábado.

De 0 a 10:

O conhecimento apresentado foi útil?

O conhecimento apresentado foi inovador?

---

**Questão 3:** Considerando a regra $X \rightarrow Y$, o fato injúria frequentemente ocorre com as vítimas de escolaridade superior.

De 0 a 10:

O conhecimento apresentado foi útil?

O conhecimento apresentado foi inovador?

---

**Questão 4:** Considerando a regra $X \rightarrow Y$, as vítimas com escolaridade superior são também as que têm uma profissão e frequentemente registram entre 1 e 3 boletins de ocorrência.

De 0 a 10:

O conhecimento apresentado foi útil?

O conhecimento apresentado foi inovador?

---

**Questão 5:** Considerando a regra $X \rightarrow Y$, as vítimas que registram ameaça recorrentemente também registram injúria. Isso se explica ao analisarmos a frase que frequentemente é proferida pelo agressor às vítimas:

"...Vou te matar (AMEAÇA) sua *algum palavrão* (INJÚRIA)..."

De 0 a 10:

O conhecimento apresentado foi útil?

O conhecimento apresentado foi inovador?

---

A Tabela 4.7 apresenta as notas atribuídas às regras pelos especialistas de domínio para o conhecimento útil extraído.

Tabela 4.7: Resultado da avaliação das regras de associação para o conhecimento útil.

| **Efetivo** | **Regra 01** | **Regra 02** | **Regra 03** | **Regra 04** | **Regra 05** |
|---|---|---|---|---|---|
| Major | 10 | 10 | 5 | 10 | 5 |
| Sargento A | 6 | 5 | 5 | 6 | 5 |
| Sargento B | 5 | 8 | 5 | 8 | 8 |
| Cabo | 8 | 8 | 6 | 10 | 6 |
| **Média** | **7.25** | **7.75** | **5.25** | **8.5** | **6.0** |

Fonte: Autoria própria.

A Tabela 4.8 apresenta as notas atribuídas às regras pelos especialistas de domínio para o conhecimento inovador extraído.

Tabela 4.8: Resultado da avaliação das regras de associação para o conhecimento inovador.

| **Efetivo** | **Regra 01** | **Regra 02** | **Regra 03** | **Regra 04** | **Regra 05** |
|---|---|---|---|---|---|
| Major | 10 | 10 | 10 | 10 | 10 |
| Sargento A | 5 | 5 | 4 | 6 | 4 |
| Sargento B | 3 | 2 | 4 | 7 | 8 |
| Cabo | 6 | 5 | 6 | 8 | 8 |
| **Média** | **6.0** | **5.5** | **6.0** | **7.75** | **7.5** |

Fonte: Autoria própria.

Conforme as Tabelas 4.7 e 4.8, a regra extraída que mais apresentou tanto conhecimento útil quanto inovador foi a de que vítimas com escolaridade superior frequentemente registram entre 1 e 3 boletins de ocorrência e são também

as que têm uma profissão. Tal análise pode ser feita por meio do resultado das suas respectivas médias: 8.5 e 7.75.

O arquivo CSV com a base de dados utilizada para a extração das regras está disponível em: `https://github.com/wesley-pm/violencia_domestica`

CAPÍTULO

# 5

# Conclusões

A quantidade de dados gerados e armazenados vem crescendo juntamente com o aumento do poder computacional de guardá-los. No Mato Grosso do Sul, os dados sobre ocorrências criminais são armazenados no Sistema Integrado de Gestão Operacional (SIGO). Atualmente, não há no Estado um departamento especializado para a análise dos dados, tampouco o emprego de técnicas de mineração de dados para a extração de conhecimento e o suporte na tomada de decisões. Todavia, devido à quantidade enorme de dados armazenados, torna-se inviável uma análise manual por busca de padrões.

Posto isso, o presente trabalho teve como objetivo principal realizar, por meio da mineração de dados, a extração de conhecimento dos crimes de violência doméstica registrados na cidade de Campo Grande (MS). Espera-se que o conhecimento obtido neste trabalho dê suporte à tomada de decisões no combate a esse tipo de crime, além de direcionar o atendimento prioritário do PROMUSE às mulheres em situação mais vulnerável, além de promover um melhor entendimento sobre os padrões de ocorrência de crimes, mais especificamente, de violência contra a mulher. Por fim, teve também como objetivo um estímulo inicial ao uso de mineração de dados na base do SIGO.

Por questões de explicabilidade e interpretabilidade dos resultados para os tomadores de decisão, foram utilizados algoritmos de mineração de dados do tipo simbólico: regras de associação e regras de classificação. Em geral, a proposta de se aplicar mineração de dados para extrair conhecimento dos crimes de violência doméstica mostrou-se eficaz: foi possível extrair conhecimento interessante e inovador tanto com as regras de associação quanto com as regras de classificação. Por exemplo, ao analisar o fato injúria, constatou-se que ele mais frequentemente ocorre com as vítimas de escolaridade superior. Já para

as regras de classificação, foi possível obter uma acurácia de 84% utilizando os algoritmos de classificação, sendo possível extrair, a partir dos dados, que as vítimas que estão há mais tempo com o agressor e têm filhos com eles estão mais propensas a permanecer no relacionamento abusivo, encontrando mais dificuldade em romper o ciclo da violência doméstica. Já as vítimas com até 23 anos de idade e pouco tempo junto do agressor conseguem romper o ciclo mais cedo. Por último, também foi possível analisar que a regra extraída com mais conhecimento tanto útil quanto inovador foi a de que vítimas com escolaridade superior são também as que têm uma profissão e frequentemente registram entre 1 e 3 boletins de ocorrência contra o mesmo agressor. Ou seja, uma vítima com alto grau de escolaridade e independente financeiramente consegue romper o ciclo da violência doméstica com mais facilidade.

Por conseguinte, fica evidente o potencial existente na descoberta de conhecimento em favor da segurança pública do Mato Grosso do Sul, visto que os resultados gerados servirão de informação privilegiada para a elaboração de estratégias públicas que visem a reduzir a violência doméstica no estado por meio de decisões baseadas em dados, e não somente por empirismo policial.

## 5.1 Principais Resultados e Contribuições

As contribuições do trabalho são:

1. Coopera para o debate sobre a aplicação de técnicas de descoberta de conhecimento em diferentes áreas da segurança pública;

2. Auxilia as políticas públicas voltadas às vítimas de violência;

3. Apoia o PROMUSE na classificação das vítimas de violência doméstica;

4. Extrai conhecimento potencialmente útil e inovador sobre o crime de violência doméstica na cidade de Campo Grande (MS).

Além disso, o presente trabalho serviu como base para o pré-projeto apresentado no XVII Prêmio Sul-mato-grossense de Inovação na Gestão Pública de Mato Grosso do Sul. Como ideia inovadora, propôs-se a aplicação de algoritmos de aprendizado de máquina para a previsão de crimes em Mato Grosso do Sul, utilizando-se dos dados de roubos, furtos, homicídios e tráfico de drogas registrados no estado nos últimos 10 anos (2011-2021) e armazenados no SIGO. Foi apresentada no projeto a possibilidade de prever um valor categórico a partir de informações contidas nos boletins de ocorrência policial referentes ao crime, tais como: o dia da semana, o período do dia, o mês do

ano, o bairro e o logradouro da ocorrência. A ideia inovadora ficou em 1º lugar [1].

## 5.2 Limitações e Trabalhos Futuros

Dentre as limitações encontradas, a maior delas foi a coleta manual dos atributos: filhos juntos, tempo juntos e quantidade de registros anteriores. Isso ocorreu devido ao fato de as informações estarem no histórico das ocorrências, onde os dados não são estruturados. Mesmo com a ferramenta de extração dos dados, ainda não há possibilidade de extrair tais informações. Isso também limitou o tamanho da base de dados, visto que ela ficou limitada ao esforço humano de abrir ocorrência por ocorrência e lançar as informações referentes aos atributos acima mencionados.

No desenvolvimento do trabalho, foram surgindo possibilidades que servem como sugestões para a realização de trabalhos futuros, conforme apresentado a seguir.

1. Adotar técnicas de mineração de texto nos históricos das ocorrências para extrair outros tipos de atributos, como armas utilizadas ou se a vítima já requereu medidas protetivas contra o agressor.

2. Utilizar aprendizado de máquina para a tarefa de regressão, tendo como objetivo prever a quantidade de ocorrências e, a partir do valor da previsão de quantos registros serão feitos pelas vítimas, priorizar o atendimento delas com base em números maiores previstos de reincidência de ocorrências.

3. Testar o emprego do aprendizado profundo na base de violência doméstica com o objetivo de melhorar a performance de classificação. Também testá-lo junto a outras fontes de dados, tais como demográficos, socioeconômicos e ambientais.

4. Houve desbalanceamento entre as classes, pretende-se aplicar em trabalhos futuros, técnicas como *oversample* para tentar solucionar o problema e, posteriormente, medir o impacto nas regras de classificação.

[1] https://www.pm.ms.gov.br/geral/tres-policiais-militares-sao-contemplados%2D-com-o-premio-sul-mato-grossense-de-inovacao/

# Referências Bibliográficas


Agrawal, R. e Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th Int. conf. very large data bases, VLDB*, páginas 487–499. Santiago, Chile. Citado na página 16.

Albani, L. (2021). Assembléia legislativa do espirito santo: Apesar dos avanços, mulheres convivem com violência. Disponível em: <https://www.al.es.gov.br/Noticia/2021/03/40569/apesar-de-avancos-mulheres-convivem-com-violencia.html>. Acesso em: 15 mar. 2023. Citado na página 7.

Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, 19(1):9. Citado na página 14.

Avila, T. P. (2021). Primary prevention policies to face domestic violence against women: Lessons from australia to brazil. *International journal for crime, justice and social democracy*, 10(4):189–203. Citado na página 5.

Azevedo, A. e Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining 2008, Amsterdam, Holanda.*, páginas 23-26. Citado na página 10.

Aziz, R. M., Hussain, A., Sharma, P., e Kumar, P. (2022). Machine learning-based soft computing regression analysis approach for crime data prediction. *Karb Int J Mod Sci*, 8(1):1–19. Citado na página 20.

Bharati, A. (2018). Crime prediction and analysis using machine learning. *International Research Journal of Engineering and Technology*, 5(9):1037–1042. Citado na página 19.

Braga, A. A. e Weisburd, D. (2010). *Policing problem places: Crime hot spots and effective prevention*. Oxford University Press on Demand. Citado na página 9.

Brayne, S. (2017). Big data surveillance: The case of policing. *American sociological review*, 82(5):977–1008. Citado na página 9.

Castro, L. N. d. e Ferrari, D. G. (2016). Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. *São Paulo: Saraiva*, 5. Citado na página 14.

Cazorla, I. M. (2003). Média aritmética: um conceito prosaico e complexo. *Anais do IX Seminário de Estatística Aplicada*, páginas 1–14. Citado na página 24.

Cerqueira, D. R. d. C., Matos, M. V. M., Martins, A. P. A., e Pinto Júnior, J. A. (2015). Avaliando a efetividade da Lei Maria da Penha. *Instituto de Pesquisa Econômica Aplicada (IPEA)*. Citado na página 5.

Corcovia, L. O. e Alves, R. S. (2019). Aprendizagem de máquina e mineração de dados: avaliação de métodos de aprendizagem. *Revista Interface Tecnológica*, 16(1):90–101. Citado na página 14.

Cunha, J. P. Z. (2019). Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. *Universidade de São Paulo*. Citado na página 15.

Dias, M. B. (2007). A lei maria da penha na justiça. *São Paulo: Revista dos Tribunais*, páginas 2–49. Citado na página 5.

Fayyad, U., Piatetsky Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37. Citado na página 12.

Ferguson, A. G. (2017). Rise of big data policing, the. In *Rise of Big Data Policing, The*. New York University Press. Citado na página 9.

Ferrari, D. G. e Silva, L. N. D. C. (2017). *Introdução a mineração de dados*. Saraiva Educaçao SA. Citado na página 16.

Garcia, C., Sanjuan, G., Puerta-Alcalde, P., Moreno-García, E., e Soriano, A. (2019). Artificial intelligence to support clinical decision-making processes. *EBioMedicine*, 46:27–29. Citado na página 1.

Goldschmidt, R. e Passos, E. (2005). *Data mining*. Gulf Professional Publishing. Citado na página 17.

Gonçalves, E. C., Freitas, A. A., e Plastino, A. (2018). A survey of genetic algorithms for multi-label classification. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, páginas 1–8. IEEE. Citado na página 14.

Han, J., Pei, J., e Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier. Citado nas páginas 1 e 12.

Hand, D., Mannila, H., e Smyth, P. (2001). Principles of data mining. 2001. *MIT Press. Sections*, 6(3):2–6. Citado na página 9.

Hassani, H., Huang, X., Silva, E. S., e Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(3):139–154. Citado na página 19.

Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M., e Sarker, I. H. (2020). Crime prediction using spatio-temporal data. In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1*, páginas 277–289. Springer. Citado na página 19.

Imielinski, T., Swami, A., e Agarwal, R. (1993). Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Conf. Management of Data*, volume 10. Citado na página 18.

Ivan, N., Ahishakiye, E., Omulo, E. O., e Taremwa, D. (2017). Crime prediction using decision tree (j48) classification algorithm. *International Journal of Computer and Information Technology*. Citado na página 19.

Katal, A., Dahiya, S., e Choudhury, T. (2022). Energy efficiency in cloud computing data center: A survey on hardware technologies. *Cluster Computing*, 25(1):675–705. Citado na página 1.

Lemos, J. L. C. (2020). Você sabe o que é CRISP-DM? Disponível em: <https://medium.com/bexs-io/você-sabe-o-que-é-crisp-dm-a3c15975bd4c>. Acesso em: 15 mar. 2023. Citado na página 11.

Lopez, C. (2021). *Data mining. The CRISP-DM methodology. The CLEM language and IBM SPSS MODELER*. United States: Lulu Press Inc. Citado na página 10.

Maimon, O. Z. (2014). *Data mining with decision trees: theory and applications*, volume 81. World scientific. Citado na página 13.

Maldonado, V. N. (2019). LGPD: Lei geral de proteção de dados comentada. *São Paulo: Revista dos Tribunais*. Citado na página 8.

Mariano, D. C. B., Marques, L. T., e Silva, M. S. (2021). *Data mining*. Porto Alegre: SAGAH. Citado nas páginas 9, 12, e 17.

Mitra, S., Pal, S. K., e Mitra, P. (2002). Data mining in soft computing framework: a survey. *IEEE transactions on neural networks*, 13(1):3–14. Citado na página 1.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrielva*, 2(1–2):1–135. Citado na página 14.

Pougy, L. G. (2010). Desafios políticos em tempos de Lei Maria da Penha. *Revista Katálysis*, 13:76–85. Citado na página 5.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier. Citado na página 19.

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda. Citado nas páginas 12 e 14.

Rossi, R. G. (2011). *Representação de coleções de documentos textuais por meio de regras de associação*. PhD thesis, Universidade de São Paulo. Citado nas páginas 14, 15, 16, e 18.

Rossi, R. G. (2015). Classificação automática de textos por meio de aprendizado de máquina baseado em redes. Citado na página 15.

Russell, S. J. e Norvig, P. (2004). *Inteligência artificial*. Elsevier. Citado na página 14.

Safat, W., Asghar, S., e Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE Access*, 9:70080–70094. Citado na página 20.

Sathyadevan, S., Devan, M., e Gangadharan, S. S. (2014). Crime analysis and prediction using data mining. In *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, páginas 406–412. IEEE. Citado na página 19.

Sherman, L. W., Gartin, P. R., e Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology*, 27(1):27–56. Citado na página 9.

Silva, E. R. G. (2011). O processo de descoberta do conhecimento como suporte à análise criminal: minerando dados da segurança pública de santa catarina. In *International Conference on Information Systems and Technology Management*, volume 8, páginas 3144–3174. Citado nas páginas 1 e 18.

Silva, L. A., Peres, S. M., e Boscarioli, C. (2017). *Introdução à mineração de dados: com aplicações em R*. Elsevier Brasil. Citado nas páginas 9, 12, 13, e 16.

Tan, P., Steinbach, M., Karpatne, A., e Kumar, V. (2019). *Introduction to Data Mining*. Global Edition. Pearson Education Limited. Citado nas páginas 12, 14, e 15.

Tan, P. N., Steinbach, M., e Kumar, V. (2016). *Introduction to data mining*. Pearson Education India. Citado na página 1.

Tsoumakas, G., Katakis, I., e Vlahavas, I. (2010). Mining multi-label data. *Data mining and knowledge discovery handbook*, páginas 667–685. Citado na página 14.

Veiga, B. C. S. (2021). Distribuição espacial da violência contra a mulher: Uma análise por geoprocessamento. *Revista Brasileira de Ciências Criminais*, 186(2021):285–319. Citado na página 6.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. Wiley. Citado na página 1.

Wang, T., C. R., Wagner, D., e Sevieri, R. (2013). Detecting patterns of crime with series finder. páginas 140–142. Citado na página 18.

Zhang, M. L. e Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837. Citado na página 14.