UNIVERSIDADE DO VALE DO TAQUARI
CURSO DE ENGENHARIA DA COMPUTAÇÃO

SISTEMATIZAÇÃO DO PROCESSO DE MINERAÇÃO E ANÁLISE DE DADOS APLICADO AO SETOR PÚBLICO

Cândida Maria Bruxel

Lajeado, junho de 2020.

Cândida Maria Bruxel

SISTEMATIZAÇÃO DO PROCESSO DE MINERAÇÃO E ANÁLISE DE DADOS APLICADO AO SETOR PÚBLICO

Monografia desenvolvida na disciplina de Trabalho de Conclusão de Curso II, do curso de Engenharia da Computação, da Universidade do Vale do Taquari - Univates, como parte da exigência para a obtenção do título de Bacharel em Engenharia da Computação.

Orientador: Prof. Dr. Marcelo de Gomensoro Malheiros

Lajeado, junho de 2020.

RESUMO

A Mineração de Dados é o processo de identificar padrões, regras, correlações e anomalias em conjuntos de dados para extrair conhecimento e ajudar gestores na tomada de decisão, além de ser uma área multidisciplinar, incluindo tecnologias de armazenamento, técnicas de Inteligência Artificial, algoritmos de Aprendizado de Máquina e estratégias de Visualização Científica. Este trabalho teve por objetivo explorar uma metodologia de Mineração e Análise dos Dados, utilizando como estudo de caso informações oriundas do setor público. Para tanto, foi estruturada uma base de dados com dados consolidados e então diversas técnicas de Mineração de Dados foram aplicadas, seguindo a metodologia Cross-Industry Standard Process for Data Mining (CRISP-DM). Além do resultado imediato de descobrir informações úteis e ainda desconhecidas, este trabalho procurou analisar ferramentas e algoritmos atualmente disponíveis, assim como refletir sobre como as informações mineradas podem ser retornadas de forma interativa para um gestor público.

Palavras-chave: Mineração de Dados. Análise de Dados. CRISP-DM. Aprendizado de Máquina. Visualização Científica. Setor Público.

ABSTRACT

Data Mining is the process of identifying patterns, rules, correlations and anomalies in data sets to extract knowledge and help decision makers, being a multidisciplinary area, which includes storage technologies, Artificial Intelligence techniques, Machine Learning methods and Scientific Visualization strategies. This work aims to explore a mining and data analysis methodology, using as a case study information from the public sector. For this, a database with consolidated data was structured and then several data mining techniques were applied, following the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. In addition to the immediate result of discovering useful and yet unknown information, this work seeks to analyze currently available tools and algorithms, as well as to reflect on how mined information can be interactively returned to a public sector manager.

Keywords: Data Mining. Data analysis. CRISP-DM. Machine Learning. Scientific visualization. Public sector.

Dedico este trabalho a minha família que sempre me apoiou durante a graduação.

AGRADECIMENTOS

Ao professor e orientador Marcelo de Gomensoro Malheiros que me ajudou no desenvolvimento do presente trabalho, e por toda orientação e dedicação.

Aos meus pais Auri e Maria Bruxel, irmã Graziela Bruxel e namorado Elton Raimondi que sempre me incentivaram a estudar e conquistar meus objetivos.

Ao meu irmão Ismael Bruxel por todo incentivo, apoio e auxílio durante a graduação.

Aos envolvidos no Projeto Lajeado pela Paz por todo apoio e disponibilidade fornecido para que esse trabalho pudesse ser desenvolvido.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
API	Application Programming Interface
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma-Separated Values
DBF	Data Base File
DBSCAN	Density Based Spatial Clustering of Applications with Noise
ETL	Extract, Transform and Load
GATE	General Architecture for Text Engineering
HTML	HyperText Markup Language
IA	Inteligência Artificial
KDD	Knowledge Discovery in Databases
MD	Mineração de Dados
PDF	Portable Document Format
SINAN	Sistema de Informação de Agravos de Notificação
SOM	Self-Organizing Map Neural Network
STHAS	Secretaria de trabalho, habitação e Assistência Social
SVM	Support Vector Machine
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

O volume de dados armazenados vem crescendo exponencialmente e em alta velocidade, principalmente devido à rápida automatização das empresas e à queda no custo do armazenamento. Desta forma, armazenar os dados em sistemas computacionais tornou-se um hábito comum e fundamental nos dias de hoje. Tipicamente esses conjuntos de dados são armazenados em *data warehouses*, bases de dados ou demais tipos de repositórios, seja de maneira centralizada ou distribuída (REZENDE, 2004).

De acordo com Dias (2002), embora as informações de um dado negócio já se encontrassem em grandes repositórios de dados, ainda havia uma grande dificuldade na descoberta de novo conhecimento, adquirido através de processamento destas informações.

Segundo Carvalho (2005), devido essa necessidade de extrair informações valiosas dos dados surge a Mineração de Dados (MD), que passa a unificar as técnicas tradicionais de análises de dados com técnicas modernas e automáticas de exploração, proporcionando a extração de relações e novos padrões que não seriam facilmente vistos a olho nu pelo ser humano, devido ao grande volume de dados envolvidos.

A principal finalidade da Mineração de Dados é a aplicação de técnicas que permitem extrair conhecimento de dados armazenados. Uma das maneiras de transformar os dados em conhecimento é empregando o uso de algoritmos

especializados, por meio de diversas ferramentas, além de processos formais, tarefas de testes e treinamento específico para se obter resultados (REZENDE et al., 2003).

Segundo Han e Kamber (2006), os métodos de Mineração de Dados dispõem de várias técnicas para a solução de problemas, diagnósticos, análises, aprendizagem, planejamento e inovação. Além disso, é uma área multidisciplinar envolvendo também em seu processo repositórios de dados, algoritmos matemáticos, técnicas estatísticas e visualização de dados.

Rezende et al. (2003), ainda complementam que as técnicas e ferramentas usadas no processo de Mineração de Dados são genéricas, possibilitando que a implementação delas ocorra normalmente independente da metodologia escolhida. A execução das técnicas exige apenas uma adaptação dos dados que deverão ser analisados.

Este trabalho propôs a aplicação e análise da metodologia CRISP-DM para efetuar Mineração de Dados. Segundo Wirth e Hipp (2000), as duas primeiras etapas desta metodologia são o entendimento dos negócios e entendimento dos dados a serem manuseados, o que permite estabelecer um problema, definir propósitos e elaborar hipóteses por meio dos dados inicialmente coletados.

O desenvolvimento do trabalho se realizou utilizando dados provenientes de um caso específico, o projeto Pacto Lajeado pela Paz da Prefeitura Municipal de Lajeado do Rio Grande do Sul (RS). Foi implementado e validado o fluxo de Mineração de Dados para alguns dados colhidos dentro do projeto. Para tanto, utilizando a metodologia CRISP-DM, foi realizado o processo de entendimento do negócio e dos dados, além da coleta e organização deles, para posteriormente aplicar técnicas de MD. Esperava-se completar o processo com o retorno para o gestor público do conhecimento especializado extraído, para validação e crítica, mas por efeito direto da pandemia, esta última etapa não foi possível.

O trabalho também analisou a melhor forma de apresentar os resultados obtidos para o usuário final, buscando uma maneira interativa para facilitar o entendimento e exploração dos resultados.

1.1 Problema de pesquisa

O processo de Mineração de Dados é capaz de auxiliar na análise de dados e extração de conhecimento útil dos dados do setor público?

1.2 Objetivos

O objetivo do presente trabalho foi realizar a execução do processo de Mineração e Análise de Dados provenientes do setor público, mapeando metodologias, técnicas e ferramentas existentes. Este trabalho visa propor um fluxo simples de organização e manipulação de dados para que análises mais complexas possam ser realizadas, podendo auxiliar na descoberta de conhecimento em bases de dados e na tomada de decisão por gestores.

Para que este objetivo fosse alcançado, os seguintes objetivos específicos foram definidos:

- Extrair e organizar os dados da área da saúde da cidade de Lajeado, implementando um fluxo de Mineração de Dados;
- Realizar análises dos dados com técnicas de Mineração de Dados, retornando os resultados preliminares para o gestor público fazer sua avaliação;
- Propor processos e ferramentas que possibilitem que as análises feitas, possam ser repetidas futuramente com novos dados.

1.3 Estrutura do trabalho

Este presente trabalho foi estruturado em seis capítulos. O primeiro capítulo trata da apresentação introdutória sobre Mineração de Dados, juntamente com o problema de pesquisa e os objetivos deste trabalho. No segundo capítulo está escrita

a fundamentação teórica, descrevendo os conceitos necessários para a compreensão deste trabalho. No terceiro capítulo são apresentados os trabalhos relacionados, em particular trabalhos acadêmicos e artigos relacionados ao uso de Mineração de Dados no setor público.

No quarto capítulo é descrita a metodologia, apresentando o tipo de pesquisa usada neste trabalho, bem como as tecnologias utilizadas e o desenvolvimento da parte técnica. No quinto capítulo são apresentados os testes e feita a análise dos resultados obtidos. Por fim, no sexto e último capítulo são apresentadas as conclusões e indicados trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo contém conceitos teóricos referentes a Inteligência Artificial, Aprendizado de Máquina, além das etapas e tarefas da Mineração de Dados. Apresentam-se também alguns procedimentos sobre como padronizar os dados de uma base de dados, a serem posteriormente utilizados em algoritmos de Mineração de Dados. Estes conceitos são essenciais para o entendimento do presente trabalho.

2.1 Inteligência Artificial

Segundo Fernandes (2008), vários pesquisadores e cientistas estão há anos estudando a inteligência humana com o objetivo de entendê-la. Embora inicialmente a área de Filosofia tivesse demonstrado mais interesse, ao longo desse período, cientistas também têm realizado inúmeras pesquisas para que se consiga reproduzir a forma humana de pensar.

Para a área da Computação, usa-se a expressão Inteligência Artificial (IA) para nomear a tentativa de simular inteligência em máquinas. Whitby (2003) define a Inteligência Artificial como o estudo sobre o comportamento inteligente, o que inclui homens, animais e máquinas, e o esforço para identificar maneiras que consigam transformar esse comportamento em qualquer tipo de obra através da engenharia.

Para Fernandes (2008), a inteligência é a capacidade de desempenhar de forma eficaz uma determinada atividade e artificial é aquilo que é desenvolvido pelo

homem. Por isso, IA é uma espécie de inteligência desenvolvida pelo próprio homem para atribuir às máquinas algum comportamento que represente a inteligência do ser humano.

Conforme Faceli et al. (2011), tradicionalmente a IA era considerada uma área de viés teórica, sendo aplicada em situações específicas e longe dos olhos do público leigo. Em geral, os problemas que necessitavam de computação eram resolvidos com codificação em determinada linguagem de programação através dos passos essenciais para a solução.

No decorrer dos anos, houve uma propagação na utilização das técnicas de computação baseadas em Inteligência Artificial para resolver problemas mais complexos. Em sua maioria, os problemas passaram a ser resolvidos computacionalmente através da obtenção de conhecimento de especialistas de determinada área, que seria então codificada em um programa de computador através de regras lógicas. Os programas criados desta forma eram assim nomeados como sistemas baseados em conhecimento ou sistemas especialistas (FACELI et al., 2011).

Com o aumento da dificuldade dos problemas que deveriam ser solucionados computacionalmente, e com o aumento da quantidade de dados gerados, novas ferramentas mais autônomas foram desenvolvidas, para que fosse possível diminuir a interação e dependência dos especialistas humanos (FACELI et al., 2011).

Segundo Rezende (2004), com os avanços da IA, ela também proporcionou que diferentes técnicas para extração ou reconhecimento de padrões sejam utilizadas, principalmente em grandes volumes de dados. Essa extração é justamente possível por meio de técnicas de manuseio de dados.

2.2 Aprendizado de Máquina

As novas técnicas desenvolvidas para resolver os problemas complexos de IA precisavam conseguir “criar” por conta própria, através de experiências passadas, ou usando uma função ou uma hipótese que fosse eficiente para solucionar o problema

em questão. Para esse processo deu-se o nome Aprendizado de Máquina (AM), ou do inglês, Machine Learning (FACELI et al., 2011).

A área de AM foca na criação de programas de computador que automaticamente melhoram seu próprio desempenho, através da experiência. Esse método de aprendizagem não é instantâneo, requerendo um processo iterativo e/ou interativo de adequação com o ambiente (HAN; KAMBER; PEI, 2012).

Coppin (2013) e Han, Kamber e Pei (2012) descrevem que há quatro métodos de Aprendizado de Máquina: aprendizado supervisionado, aprendizado não-supervisionado, aprendizado semi-supervisionado e aprendizado por reforço. Porém os mais utilizados e que estão diretamente relacionados à Mineração de Dados, são dois, descritos a seguir:

- Aprendizado supervisionado: é embasado em um grupo de registros pelos quais as saídas corretas são conhecidas, para que seja avaliado se o sistema teve o comportamento esperado.

- Aprendizado não-supervisionado: é embasado apenas nos registros da base, cujas saídas são desconhecidas. O algoritmo precisa aprender a rotular ou categorizar os registros.

2.3 Mineração de Dados

Para Castro e Ferrari (2016), mineração é como a remoção de minerais preciosos de uma mina. A particularidade destes materiais é que não podem ser produzidos de forma artificial, sendo ainda de fontes desconhecidas. Este processo exige acesso à mina, utilização de ferramentas apropriadas para mineração, extração do minério e preparo para seguir para a comercialização.

Desta forma, a Mineração de Dados (ou em inglês, Data Mining), recebe esta comparação ao processo de mineração de minerais, uma vez que explora uma base de dados (mina) utilizando algoritmos (ferramentas) apropriadas para obter o conhecimento (minerais preciosos) (CASTRO; FERRARI, 2016).

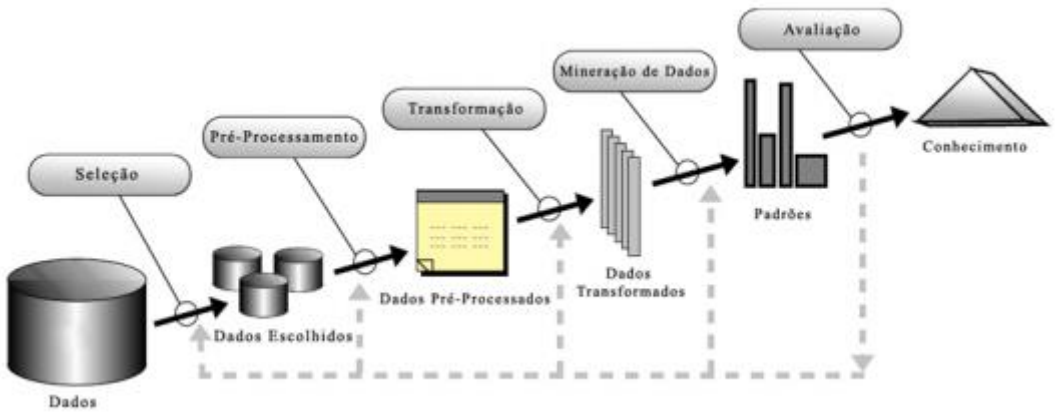
Segundo Berry e Gordon (1997), a Mineração de Dados é um processo de investigação e análise de um grande volume de dados para descobrir padrões e regras que representem algo significativo. Conforme Castro e Ferrari (2016), os dados são símbolos sem estrutura e significado, enquanto a informação extraída é o que acrescenta serventia e valor aos dados. Contudo, o conhecimento sobre essas informações é o que possibilita a tomada de decisão.

2.3.1 Knowledge Discovery in Databases

Tradicionalmente, a transformação dos dados em conhecimento, consistia em um processo exaustivo e manual realizado por especialistas, que criavam um relatório para ser analisado. Contudo, esse processo manual tornou-se inatingível com grandes volumes de dados. Assim, surge o KDD (Knowledge Discovery in Databases), com o objetivo de resolver o problema sobre uma vasta quantidade de dados. O KDD caracteriza-se em um processo complexo de descoberta de novos padrões reais, úteis e inteligíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Segundo Camilo e Silva (2009), há inúmeras definições relacionadas ao KDD e à Mineração de Dados. Existe até quem diga que são sinônimos. Já Fayyad, Piatetsky-Shapiro e Smyth (1996), definem KDD como o processo de descoberta de conhecimento, e Mineração de Dados apenas como uma das atividades mais amplas deste processo. Na Figura 1 podem-se ver as etapas do processo de KDD.

Figura 1 – Etapas do processo de KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996).

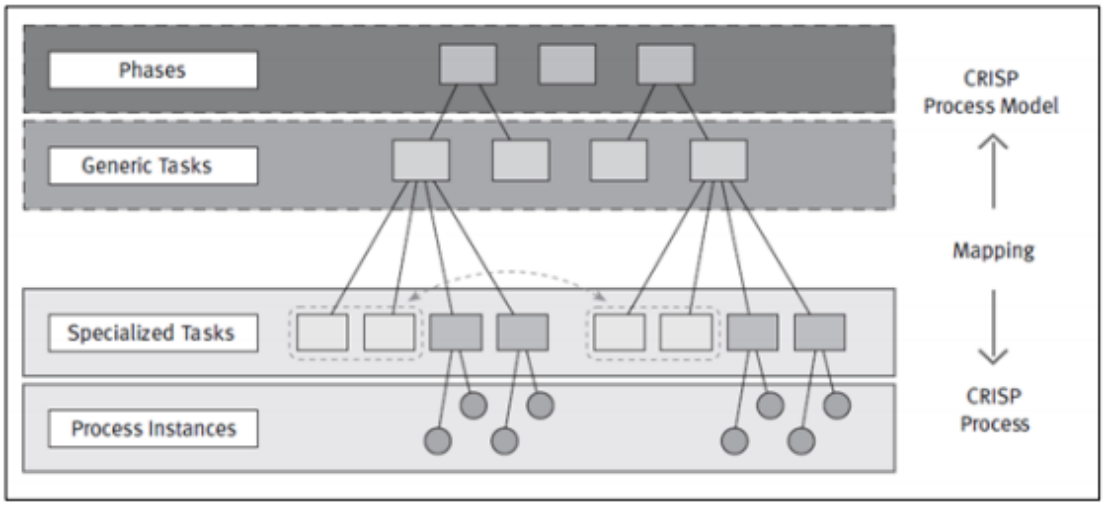
Embora o KDD seja tradicionalmente muito conhecido e utilizado, neste trabalho foi utilizada a metodologia CRISP-DM. Esta é mais recente e vem sendo considerada a metodologia de maior aceitação, uma vez que o foco do CRISP-DM é identificar o problema negocial que se deseja resolver.

2.3.2 Cross-Industry Standard Process for Data Mining

Segundo Uber (2004), o CRISP-DM foi concebido em 1996 com o objetivo de possibilitar a uniformização de técnicas e conceitos para auxiliar na busca de conhecimentos específicos, facilitando a tomada de decisão. Os criadores desenvolveram esta metodologia para apoiar os responsáveis pelo processo de planejamento e execução de Mineração de Dados, incluindo o detalhamento do processo até a visualização dos resultados. Neste grupo de criadores, havia três empresas pioneiras na área de Mineração de Dados: a DaimlerChrysler (montadora de carros), a SPSS (empresa de softwares estatísticos) e a NCR (focada em *data warehousing*).

Esta metodologia de Mineração de Dados foi elaborada como um modelo de processo hierárquico, formado por um grupo de tarefas em quatro camadas abstratas: fases, tarefas genéricas, tarefas específicas e instâncias de processos, tal como mostrado na Figura 2.

Figura 2 - Modelo de processo hierárquico do CRISP-DM

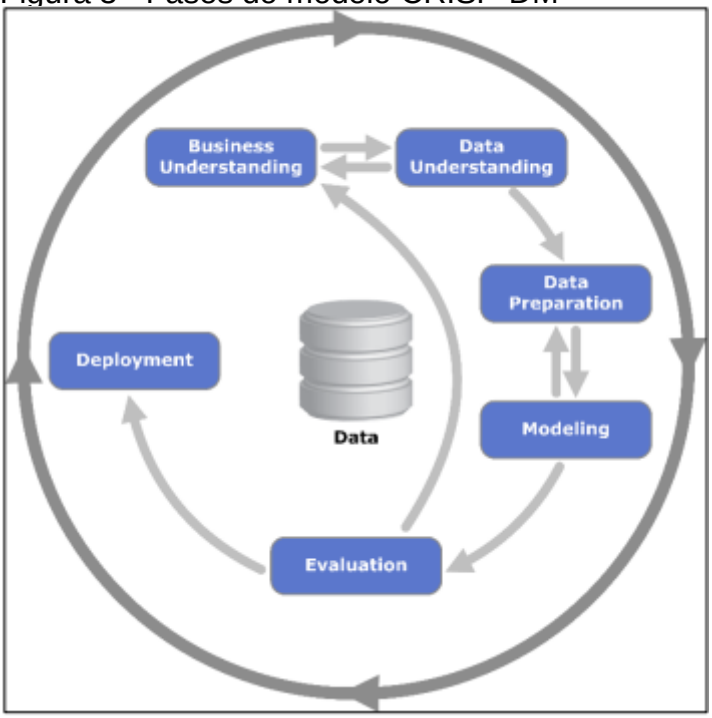


Fonte: Chapman et al. (2000).

De acordo com Chapman et al. (2000), na primeira camada várias fases são organizadas para o processo de Mineração de Dados, sendo que cada uma destas fases compõem um grupo de tarefas genéricas da segunda camada. A segunda camada tende ser mais genérica, para que se consiga abranger o máximo de situações possíveis de Mineração de Dados. Já na terceira camada, de tarefas especializadas, são definidas como serão executadas as ações da segunda camada. Na quarta camada são documentadas as ações, decisões e resultados da Mineração de Dados, sendo organizadas conforme as tarefas definidas nas camadas mais altas.

Conforme Camilo e Silva (2009), o CRISP-DM disponibiliza uma visão global do ciclo de vida de um projeto de Mineração de Dados. Nesta metodologia, o ciclo de vida possui seis fases: Entendimento dos Negócios (Business Understanding), Entendimento dos Dados (Data Understanding), Preparação dos Dados (Data Preparation), Modelagem (Modeling), Avaliação (Evaluation) e Implantação (Deployment). A Figura 3 ilustra essas fases.

Figura 3 - Fases do modelo CRISP-DM



Fonte: Chapman et al. (2000).

2.4 Fases do CRISP-DM

As fases do CRISP-DM não são necessariamente executadas em sequência, uma vez que durante o processo podem ocorrer alterações e ser modificada a sequência, como mostram as setas na Figura 3. As setas externas do ciclo sinalizam a sequência padrão da Mineração de Dados e as internas simbolizam as possíveis alterações. O próprio resultado de cada fase poderá determinar a próxima fase a ser executada (CHAPMAN et al., 2000).

2.4.1 Entendimento dos Negócios

Essa fase, de acordo com Azevedo e Santos (2008) e Silva (2002), dedica-se à compreensão dos requisitos e propósitos do projeto, sob uma visão de negócios e, a partir deste aprendizado, é possível definir um problema de mineração.

2.4.2 Entendimento dos Dados

A fase de entendimento, inicia-se com a coleta básica de dados e continua com o processo de familiarização dos dados, de descobrimento da qualidade dos dados e de identificação de grupos peculiares, com o intuito de desenvolver hipóteses sobre informações ocultas (CHAPMAN et al., 2000).

Esse procedimento inicial dos dados busca identificar as informações com as quais serão feitos os trabalhos, mapeando suas fontes, seu processo de obtenção e detectando problemas. Descreve-se também a maneira na qual os dados são obtidos, informando seu formato, volume, significado e demais informações interessantes (SILVA, 2002).

2.4.3 Preparação dos Dados

Camilo e Silva (2009) relatam que devido às diversas origens dos dados, esta fase visa construir uma base de dados final, que será aquela utilizada pelas ferramentas de mineração. Esta fase inclui as tarefas de seleção de registros e atributos, assim como transformação e limpeza da base de dados, sendo possível que estas tarefas sejam executadas repetidas vezes.

2.4.4 Modelagem

Nesta fase, vários algoritmos de Mineração de Dados são selecionados e aplicados. A escolha destas técnicas é realizada conforme o propósito desejado. Devido algumas técnicas terem requisitos particulares de formatação de dados, pode ser necessário voltar para a fase anterior (CAMILO; SILVA, 2009).

A modelagem é a etapa central da mineração, incluindo escolha, parametrização e execução de algoritmos utilizando a base de dados final, tendo o objetivo de criar um ou mais modelos (SILVA, 2002).

2.4.5 Avaliação

Azevedo e Santos (2008) explicam que nessa fase os resultados obtidos são analisados com mais detalhes, quando especialistas do negócio são necessários. Ferramentas gráficas facilitam a visualização e análise destes resultados, sendo que o ponto crucial desta fase é identificar se algum problema de negócio ainda não foi detectado. Ainda é feito o acompanhamento para verificar que o processo de fato atingirá o propósito do negócio.

2.4.6 Implantação

Está é a última fase do ciclo, ainda que normalmente não seja o fim do projeto. O conhecimento obtido deverá ser apresentado para o cliente de maneira que ele compreenda a informação (CHAPMAN et al., 2000).

2.5 Padronização dos dados

Conforme Faceli et al. (2011), as condições dos dados influenciam diretamente o desempenho dos algoritmos escolhidos para a extração de conhecimento da base de dados. Os dados podem estar em diferentes formatos, sejam qualitativos (quando estão em forma simbólica ou categórica), ou quantitativos (quando estão em forma numérica, para representar quantidades). Eles também podem estar coerentes ou conter imperfeições e ruídos, com valores inconsistentes, incorretos, duplicados ou até mesmo ausentes. A base ainda pode apresentar muitos ou poucos registros, sendo que estes podem ou não conter muitos atributos.

Ainda segundo Faceli et al. (2011), as técnicas de processamento de dados são constantemente usadas para eliminar ou ao menos minimizar os problemas citados anteriormente, tendo como consequência a melhoria na qualidade dos dados. Esse processo facilita o uso dos algoritmos de mineração, o que possibilita a construção de modelos mais reais e coerentes. Além disso, essas técnicas também tornam a base mais adequada aos algoritmos escolhidos, para o caso deste trabalhar somente com dados numéricos, por exemplo.

Em relação à estrutura dos dados, conforme Castro e Ferrari (2016), eles podem ser:

- Estruturados: quando a estrutura do modelo de dados está completa e os dados possuem campos fixos na base. Dependem da construção de um modelo de dados, em outras palavras, a delimitação dos registros junto com suas propriedades e relações. O modelo define todos os tipos de dados que serão registrados, acessados e processados.

- Semiestruturados: quando a estrutura do modelo de dados não está completa, mas, no entanto, também não está totalmente desestruturada. Nessa estrutura geralmente são utilizados marcadores para identificação de determinados elementos, mas ela não é sólida.

- Não estruturados: quando não possui um modelo de dados, não estando classificados de uma forma predefinida.

Um método utilizado para realizar a padronização dos dados é chamada de Extract, Transform and Load (ETL), e é nessa etapa que os dados são preparados antes de serem carregados em uma ferramenta para análise.

Segundo Kimball e Caserta (2004), o ETL pretende atuar com todo tipo de exportação, transformação e importação dos dados. Na etapa de extração (Extraction) os dados são extraídos dos sistemas de origem. Já na etapa de transformação (Transform) é realizada a limpeza dos dados. Por último, na etapa de carregamento (Load), uma vez tendo a base de dados pronta, pode-se importar ela para a ferramenta utilizada para análise

O método de ETL é muito utilizado para o processo de carga em um *data warehouse*, sendo tradicionalmente aplicado em projetos de Business Intelligence. No entanto, para este trabalho esse método é indiretamente seguido, visto que a padronização de dados também envolve etapas de extração, transformação e importação nas ferramentas que serão realizadas.

2.5.1 Integração de dados

Segundo Carvalho (2005), os dados a serem utilizados para a Mineração de Dados podem estar armazenados em diferentes bases de dados, sendo que estas bases necessitam ser integradas em uma única base, antes da utilização das técnicas de mineração. Desta forma, na união das bases, é imprescindível a identificação dos objetos existentes em cada conjunto de dados a ser integrado.

Castro e Ferrari (2016) destacam três pontos que devem ser observados durante a união das bases:

- Redundância: ocorre quando um objeto/atributo derivar de um ou mais objetos/atributos da base. Para detectar dados redundantes a análise de correlação pode ser aplicada, calculando o quanto dois atributos estão numericamente correlacionados.

- Duplicidade: quando entidades muito semelhantes estão definidas em bases diferentes com nomes e atributos diferentes, necessitando assim relacioná-las.

- Conflitos: ocorrem quando os valores dos dados aparecem diferentes em fonte de dados distintas. Esse conflito pode ser resultado de diferentes formas de representações destes dados (quilômetros, metros e milhas, por exemplo).

2.5.2 Limpeza dos dados

Dados reais possuem grandes possibilidades de serem incompletos, inconsistentes e diferentes. As etapas de limpeza de dados buscam completar valores ausentes, tratar dados ruidosos, ao mesmo tempo que encontram valores discrepantes (*outliers*) e corrigem inconsistência nos dados (HAN; KAMBER; PEI, 2012).

Referente aos valores ausentes (*missing values*), Castro e Ferrari (2016) relatam que em uma base de dados podem existir diversos registros com atributos que não apresentam valores armazenados ou que possuem apenas um símbolo (por exemplo, “?”), e que ainda assim podem ser necessários para a etapa de Mineração de Dados.

Segundo Han, Kamber e Pei (2012), os métodos mais comuns para preencher valores ausentes são:

- Ignorar o registro, ou seja, excluir da base (ignorar) todos registros que possuem um ou mais valores ausentes;

- Imputar o valor manualmente, onde é inserido (respeitando o domínio de cada atributo) um valor de maneira empírica para cada dado ausente;

- Usar uma constante global para preencher os valores ausentes, onde se insere (respeitando o domínio de cada atributo) em todos os valores ausentes de determinado atributo, um único valor constante;

- Imputação do tipo hot-deck, onde encontra-se o registro com valor observado mais similar com o registro com valor ausente em relação com os demais atributos, inserindo o valor do correspondente no valor ausente;

- Imputação de acordo com a última observação, onde ordena-se a base de dados conforme um ou mais de seus atributos. Assim na sequência o algoritmo buscará os valores ausentes e usará o valor do registro anterior para inserir no lugar do dado faltante;
- Usar a média ou moda, onde os dados faltantes de cada atributo serão substituídos pela moda (no caso de dados qualitativos) ou média (no caso de dados quantitativos) dos valores do atributo.

Dados ruidosos, segundo Faceli et al. (2011), são dados fora do padrão que aparentam não pertencer ao arranjo que concebeu os dados analisados. Ruído é uma variação acentuada ou um erro aleatório na mensuração de uma variável.

Conforme Han, Kamber e Pei (2012), existem alguns métodos para corrigir esses valores:

- O encaixotamento (*binning*), em que primeiramente ordenam-se os valores do atributo, assim sendo possível usar a abordagem de vizinhança entre eles. Feito isso, os valores são separados em grupos (*bins*) contendo cada grupo a mesma quantidade de elementos. Para cada grupo, medidas para ajustar os valores são tomadas, como mediana, média aritmética ou um valor de limite;

- Agrupamento (*clustering*), onde valores fora do padrão podem ser descobertos, quando os elementos semelhantes são agrupados, e consequentemente os elementos que estão fora do normal podem ser apontados como ruidosos;

- Aproximação, quando nota-se que um atributo de determinado objeto com dados ruidosos tende a se distanciar dos outros objetos de sua classe. Para suavizar esta distância, técnicas baseadas em distâncias examinam a classe em que estão os objetos mais próximos;

- Regressão ou classificação, quando utiliza-se de uma função de regressão linear para estimar o valor correto de um valor com ruído. Caso o valor seja simbólico, usa-se a técnica de classificação.

Dados inconsistentes, por outro lado, são aqueles que apresentam valores incompatíveis em seus atributos. Essas inconsistências são encontradas quando vínculos conhecidos entre os atributos são violados. Podem ocorrer devido à integração de dados de diferente fontes ou ruídos na origem dos dados (FACELI et al., 2011).

2.5.3 Redução dos dados

Devido ao enorme volume de dados em determinadas bases de dados, métodos de redução de dados acabam sendo aplicadas para que se tenha uma representação menor da base. Contudo, a integridade dos dados originais deve ser mantida, para que a mineração com os dados reduzidos possa, ao mesmo tempo, ser mais eficiente e capaz de reproduzir os mesmos resultados analíticos (HAN; KAMBER; PEI, 2012).

Alguns métodos de redução que se destacam, segundo Castro e Ferrari (2016), são:

- Seleção de atributos: consiste em diminuir a dimensionalidade, ou seja, remover atributos irrelevantes ou pouco relevantes.

- Compreensão de atributos: consiste na diminuição da dimensionalidade através de transformação de dados ou de algoritmos de codificação.

- Redução no número de dados: dados são removidos, substituídos ou estimados por uma representação mais simples. Por exemplo, métodos paramétricos podem ser utilizados para estimar os dados, de modo que apenas os parâmetros sejam armazenados. Também pode-se usar métodos não paramétricos para armazenar representações reduzidas dos dados, como histogramas, *clusters* e amostras.

- Discretização: consiste em aumentar o intervalo entre os atributos, para diminuir a quantidade dos mesmos.

2.5.4 Discretização

De acordo com Castro e Ferrari (2016), há algoritmos que são incapazes de serem aplicados com atributos numéricos. Nesses casos é preciso efetuar a discretização dos mesmos. Esta pode ser executada por meio de métodos de análise de histograma, de encaixotamento ou através da distribuição dos valores em intervalos, de forma que o valor atribuído a cada intervalo represente a média ou mediana dos valores originais.

2.5.5 Transformação dos dados

Várias técnicas de Mineração de Dados são limitadas a trabalhar somente com valores numéricos, enquanto outras somente com valores categóricos. Quando isto acontece, é preciso realizar a transformação dos valores numéricos em categóricos, ou vice-versa (CAMILO; SILVA, 2009).

Não há critérios para esta transformação e diferentes técnicas podem ser empregadas, conforme a necessidade. Algumas das técnicas aplicáveis são: suavização (remoção de valores incorretos dos dados), agrupamento (união de valores em faixas sumarizadas), generalização (transformação de valores específicos

em valores genéricos), normalização (organização das variáveis em uma só escala) e criação de novos atributos (construídos e adicionados a partir de outros já disponíveis) (CAMILO; SILVA, 2009; HAN; KAMBER; PEI, 2012).

2.6 Tarefas da Mineração de Dados

Para de Amo (2004), uma tarefa consiste em especificar que tipo de informação está sendo procurado nos dados. A definição do uso de determinada tarefa é realizada conforme os propósitos desejados para a solução a ser descoberta, estando relacionada à escolha do algoritmo utilizado, que é responsabilidade do especialista de dados.

Castro e Ferrari (2016) conceituam cada tarefa conforme descrito a seguir:

- Predição (classificação ou estimação): relacionada à aplicação de um modelo para avaliação de um objeto não rotulado ou para prever o valor de um ou vários atributos de determinado objeto. Pode ser classificada em dois tipos: classificação, quando utilizada para indicar variáveis discretas; e regressão, para indicar variáveis contínuas.

- Análise descritiva de dados: torna possível mensurar, explorar e descrever particularidades representativas. É realizada no início do processo de Mineração de Dados, possibilitando a sumarização e entendimento dos objetos da base e seus atributos.

- Análise de grupos: visa encontrar e aproximar os registros semelhantes em grupos. O agrupamento, também conhecido como *cluster*, é um conjunto de objetos semelhantes entre si, no entanto diferentes do restante dos grupos. Nesta tarefa os objetos não necessitam ser previamente categorizados.

- Associação: tem o propósito de identificar relações entre os atributos e não entre os objetos.

- Detecção de anomalias: tem como objetivo encontrar registros distintos do comportamento dos demais registros.

Tanto Camilo e Silva (2009), como Larose e Lerose (2014) nomeiam estas tarefas de forma similar: descrição, classificação, estimativa, predição, agrupamento e associação. Embora a nomenclatura seja diferente, os conceitos são os mesmos.

.

3 TRABALHOS RELACIONADOS

Este capítulo descreve trabalhos, cujo foco é a aplicação de técnicas de Mineração de Dados em bases de dados da área de segurança pública. Apesar de alguns destes trabalhos utilizarem algoritmos que não serão utilizados na presente monografia, ainda assim são relevantes pelo uso do processo de MD sobre dados do setor público.

Estes trabalhos foram selecionados através de plataformas que permitem pesquisar por artigos científicos e acadêmicos, utilizando palavras-chaves. Então serão apresentados a seguir, os trabalhos relacionados às palavras-chaves: Mineração de Dados e Setor Público. Que atingiram os critérios de maior relevância com o presente trabalho e produzidos recentemente.

O trabalho descrito por Neto (2017) tem como objetivo demonstrar como o uso de Mineração de Dados pode auxiliar na análise de ocorrências criminais. Este, busca identificar na cidade de Fortaleza e região metropolitana quais regiões possuem o maior índice de criminalidade, assim como analisar os principais crimes nos bairros que integram essas regiões. Desta forma, órgãos responsáveis podem identificar melhores maneiras de como atuar no combate ao crime.

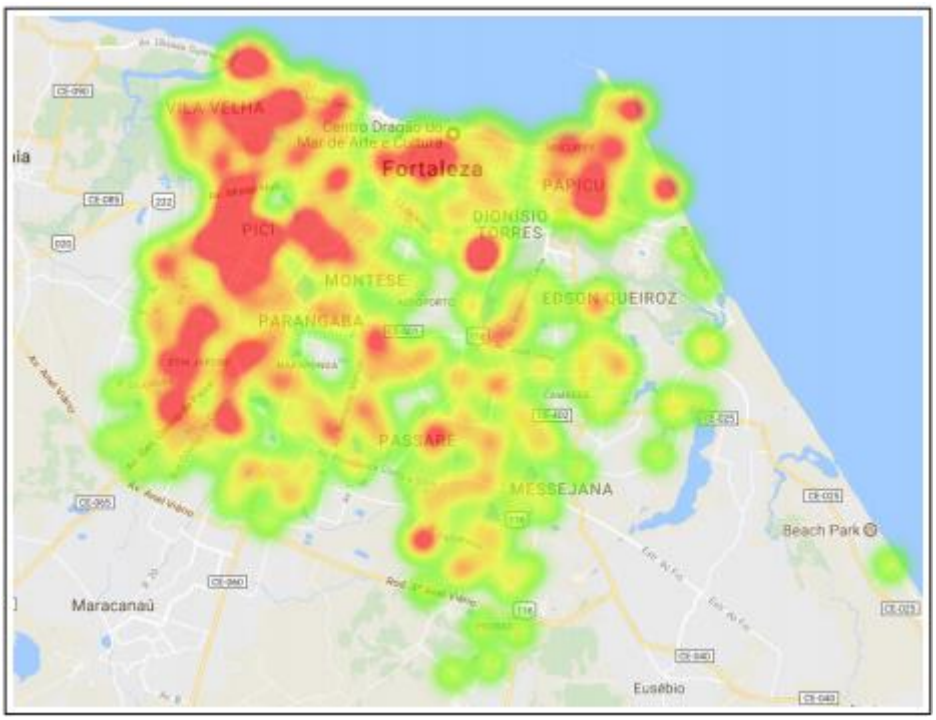
Esse estudo utilizou técnicas de descoberta de conhecimento em bancos de dados, extração de entidades e mineração. Os dados foram obtidos através do portal da Secretaria de Segurança Pública e Defesa Social do Ceará, onde estes dados foram agrupados por mês e selecionados somente os dados de janeiro até maio de 2017.

Para a identificação de regiões com o índice mais elevado de criminalidade, foi escolhida a técnica de clusterização chamada DBSCAN (Density Based Spatial Clustering of Applications with Noise). Tal algoritmo é do tipo *cluster*, tendo como propósito agrupar registros por região. Ou seja, permite agrupar pontos próximos embasado na condensação de cada registro.

Para que os dados pudessem ser minerados neste algoritmo, foi executado um algoritmo na linguagem Python, que lê a tabela do local do crime no banco de dados e retorna as respectivas informações de latitude e longitude, com o auxílio da interface do Google Maps Application Programming Interface (API).

Com essas informações também foram montados mapas de calor da cidade, onde foi possível visualizar superficialmente as regiões com maior intensidade de criminalidade. As imagens destes mapas foram criadas através da API do Google chamada Maps JavaScript API2. A Figura 4, apresenta o mapa de calor das ocorrências criminais de janeiro de 2017 da cidade.

Figura 4 - Mapa de calor das ocorrências criminais de janeiro



Fonte: Neto (2017).

Após este processo de conversão e clusterização foi realizada a análise dos resultados para cada mês. Como conclusão, o estudo conseguiu constatar a região

com mais ocorrências registradas, assim como os bairros em específico e qual a arma que mais foi utilizada para esses crimes.

Já no trabalho descrito por Keyvanpour, Javideh e Ebrahimi (2011), o objetivo foi desenvolver um estudo referente à criminologia, no qual foi utilizada a Mineração de Dados para detectar e explorar casos criminais, associando crimes com criminosos e também criminosos com crimes. O principal propósito deste estudo foi implementar uma maneira eficaz de investigação de crimes, de forma a auxiliar na tarefa de análise.

Para a realização do estudo, embora já houvesse uma base de dados disponível para análise, os autores se basearam na descrição narrativa em forma de texto simples dos crimes, realizada pelos próprios policiais. Para a utilização destas informações, foi aplicado o método de pesquisa léxico, que é um método muito comum na extração de entidades de relatórios narrativos.

Assim, juntamente com um especialista em crimes, foram definidas palavras e expressões importantes nestes textos que serviram como uma espécie de dicionário, para que fosse possível implementar uma ferramenta de pesquisa simples, capaz de encontrar nas narrativas palavras deste dicionário. Esta ferramenta foi criada para evitar a inserção manual das informações em um banco de dados estruturado.

Após esse processo de geração da base de dados, foi utilizado o algoritmo Self-Organizing Map Neural Network (SOM), capaz de encontrar a relação entre os crimes registrados. Para comparar crimes foram levados dois pontos em consideração. O primeiro ponto é quando um novo crime acontece, com o objetivo de encontrar possíveis criminosos responsáveis pelo mesmo. O segundo ponto é quando um criminoso é preso e se pretende descobrir quais outros crimes semelhantes, não resolvidos, podem ter sido cometidos por este criminoso.

A pesquisa se atribui principalmente em crimes de roubo e leva em consideração quatro variáveis: tipos de locais, interação do criminoso, métodos de abordagem e armas usadas. Para a correspondência dos crimes ser garantida, ainda foi utilizada uma rede neural Multi-Layer Perceptron (MLP) para cada variável criminal. Essa abordagem foi implementada com três camadas de neurônios: saída, entrada e uma camada oculta.

Em um terceiro estudo, relatado por Jayaweera et al. (2015), foi desenvolvido um estudo de análise criminal com dados da Internet, com base em reportagens *online* de jornais ingleses do Sri Lanka. A motivação do trabalho deveu-se ao fato dos dados estarem registrados em papéis, forçando uma análise manual e laboriosa pelos policiais. O estudo teve o propósito de auxiliar na aplicação da lei, tentando interligar incidentes criminais entre as agências policiais.

Para o desenvolvimento desse estudo, os autores utilizaram o rastreador Crawler4j, que é uma ferramenta programada para buscar informações específicas na Internet. Ela foi então utilizada em buscas pelos artigos criminais publicados em jornais, sendo então armazenados em um banco de dados. Estes foram classificados categoricamente como criminal e não-criminal, utilizando o algoritmo Support Vector Machine (SVM).

Na extração de entidades e processamento de texto, foram retirados pequenos trechos dos artigos, que foram classificados em categorias pré-definidas, usando para isso o framework General Architecture for Text Engineering (GATE). Como o armazenamento das reportagens foi feito de modo automático a cada descoberta na Internet, foi necessário tratar textos duplicados. Esta detecção de duplicatas foi calculada a partir das entidades extraídas através da técnica de SimHash, utilizando o valor resultante deste cálculo para aplicar a distância de Hamming.

Três testes foram realizados e analisados. Primeiro, a detecção de pontos quentes, para identificar áreas com maior índice de criminalidade, para que policiais pudessem interagir aumentando o número de patrulha, por exemplo e que turistas pudessem verificar se a região escolhida para conhecer é adequada. Também foram comparados crimes em um determinado período, para que se pudesse identificar que tipo de crime está precisando de mais atenção, para que medidas pontuais sejam tomadas.

Além disso, também foi disponibilizada uma visualização de padrões de crimes por meio de um gráfico de série temporal, gerado para representar as mudanças na frequência dos tipos de crime, assim auxiliando os agentes da lei e identificando a relação do tempo com mudanças graduais das frequências de crimes.

Outro estudo realizado foi relatado por Gandhi e Sharma (2017), concentrando-se em realizar uma análise preditiva para poder prever atos criminosos contra mulheres em distritos e estados da Índia, além de minerar dados e identificar padrões destes crimes.

A Índia possui um banco de dados chamado Open Government Data Platform, que na época continha informações de ocorrências criminais de 2001 a 2013. Um dos objetivos foi prever tendências de crimes até o ano de 2019. No trabalho foram considerados crimes diretamente relacionados às mulheres, tais como sequestro, estupro, agressão familiar ou feminicídio. O estudo relata ter auxiliado os órgãos de segurança pública a tomar medidas decisivas para promover ações contra tais crimes.

Para a realização dos testes foram levados em consideração alguns pontos, como qual estado tem maior números de crimes cometido, quais os crimes mais cometidos, qual estado e qual cidade têm o maior índice, qual crime teve maior número de prisões efetuadas, e qual o grupo etário.

Para a aplicação da análise foi executado o algoritmo de regressão linear e função de erro para cada cidade e estado. Foi criada uma tabela de crimes com maior índice de ocorrências, com o objetivo de prever tendências até 2019. Então foi ajustado um modelo de regressão linear usando mínimos quadrados ordinários, implementado na biblioteca Scikit-learn do Python.

Os autores constataram que o crime mais relatado é o de crueldade do marido com a esposa. Também foi identificado que conforme mais crimes eram registrados por parte das mulheres referente aos abusos domésticos no decorrer dos anos, o número de agressões e sequestros também aumentavam.

Posteriormente, executando um algoritmo de agrupamento K-Means, os autores identificaram que no ano de 2013 o número de estupros foi superior aos demais anos, o que suspeitam ter relação com os protestos femininos no ano de 2012. O trabalho não divulgou qual cidade e estado com mais ocorrência registradas, nem quais as faixas etárias do agressor e da vítima.

O último estudo, realizado por Braz et al. (2009), teve como objetivo aplicar técnicas de mineração sobre o banco de dados da Polícia Militar do Estado de

Alagoas, o SISGOP, onde se encontram armazenados boletins de ocorrência de determinadas cidades de Alagoas. O propósito foi de identificar as localidades acometidas por crimes, de determinar perfis de criminosos e vítimas, além de descobrir os dias da semana em que mais acontecem crimes.

Como os processos de análise de crimes são complexos e demorados, a ideia do estudo foi agilizar este procedimento, a ponto de que os policiais tenham mais tempo para atender à maioria das ocorrências.

Para isso os autores implementaram a metodologia CRISP-DM, impondo compreensão do negócio, planejamento detalhado e avaliação do processo de forma objetiva a cada fase e de seus problemas. Para a compreensão do negócio, os autores identificaram que nas ocorrências criminais havia dados dos anos de 2008 e 2009, e que estes eram referentes ao crime, perfil da vítima e do delinquente, incluindo ainda bairro e cidade do ocorrido.

Para a execução dos testes foi utilizada a ferramenta Waikato Environment for Knowledge Analysis (WEKA), que implementa diversas técnicas de Mineração de Dados. Usando o módulo Explorer desta ferramenta, foram feitos testes processando a base de dados com o algoritmo K-Means, para agrupar conforme perfil da vítima e do criminoso, a fim de encontrar associações.

O resultado dessa análise foi descrito como satisfatório, sendo possível identificar o perfil de vítimas mais afetado, no caso: mulheres, vítimas de roubo e ameaça no Jacintinho (bairro de Maceió) entre as 14 horas e 20 horas no dia de domingo. Já no resultado do teste de perfil de criminoso, houve muitos casos que chamaram a atenção, mas o que se destaca é o da cidade de Arapiraca, onde o perfil é masculino e o crime foi de porte de arma.

No Quadro 1, é feita uma breve comparação entre os trabalhos relacionados com o presente trabalho, envolvendo a área de origem dos dados, a metodologia utilizada, o tipo de análise feita e também os principais algoritmos utilizados.

Quadro 1 - Trabalhos relacionados

Trabalho	Criminalidade	CRISP-DM	Mineração de dados	Técnica
Neto (2017)	X		X	DBSCAN
Keyvanpour, Javideh e Ebrahimi (2011)	X		X	SOM
Jayaweera et al. (2015)	X		X	SVM
Gandhi e Sharma (2017)	X		X	K-MEANS
Braz et al. (2009)	X	X	X	K-MEANS
Presente trabalho		X	X	K-MEANS; APRIORI

Fonte: Da autora (2020).

O propósito deste trabalho é um pouco diferente das abordagens dos trabalhos relacionados, visto que a origem dos dados é a violência e não a criminalidade. No entanto todos estão relacionados a Análise e Mineração de Dados voltados a dados públicos.

4 MATERIAIS E MÉTODOS

O presente trabalho apresenta-se como um estudo de caso bibliográfico, exploratório e descritivo, de abordagem qualitativa e quantitativa.

Os objetivos foram atingidos por meio de uma pesquisa descritiva e exploratória. Segundo Prodanov e Freitas (2013), a pesquisa descritiva tem como objetivo retratar e escrever os eventos sem alterá-los, relatando particularidades de determinados grupos analisados ou o relacionamento entre ocorrências. Kauark, Manhães e Medeiros (2010) ainda complementam que este tipo de pesquisa utiliza técnicas padronizadas para a coleta de informações.

Conforme Cervo, Bervian e da Silva (2007), a pesquisa exploratória é vista por vários autores como praticamente científica, uma vez que é aplicada como primeiro passo para o procedimento de pesquisa e proporciona um auxílio para o procedimento de elaboração de hipóteses cruciais para pesquisas seguintes. Para Kauark, Manhães e Medeiros (2010), ela induz o pesquisador a ter um conhecimento mais profundo do assunto.

Em resumo, o presente trabalho pode ser considerado descritivo e exploratório, visto que foi analisado o processo de Mineração de Dados, utilizando suas tarefas e técnicas, e também explorando os dados de um caso em particular, no qual conhecimento útil foi extraído deste.

Quanto à natureza da abordagem, Prodanov e Freitas (2013) descrevem que o método qualitativo necessita somente do que se foi observado e o que isto representa,

sem a necessidade de dados numéricos ou outros recursos. Kauark, Manhães e Medeiros (2010), complementam que podemos levar em conta que haverá uma ligação entre a realidade e o indivíduo, e as respostas serão descritivas.

Já no método quantitativo, leva-se em consideração tudo que é mensurável, isto é, qualquer informação que pode ser retratada em números ou informações que possam ser classificadas. É necessária a criação de suposições e a categorização do vínculo destes fatos para se obter uma conclusão concreta, sem que haja discordância na compreensão do resultado (PRODANOV; FREITAS, 2013).

Para o presente trabalho foi utilizada tanto a pesquisa quantitativa, como a qualitativa. Quantitativa, pois foram utilizadas de medidas estatísticas para análise dos dados da Prefeitura de Lajeado, aplicando nestes dados, técnicas de Mineração de Dados para obtenção de indicadores de interesse. Qualitativa, pelo fato de que análises e reflexões foram realizadas, tanto sobre os dados estatísticos, como os resultados dos algoritmos de MD.

A pesquisa bibliográfica consiste na preparação de um estudo com base em dados já divulgados, integrado de literatura física ou *online*, com o propósito de que o explorador tenha proximidade total com o tema em questão (PRODANOV; FREITAS, 2013).

Já o estudo de caso, consiste em desenvolver uma investigação abrangente e meticulosa de um ou poucos elementos. Engloba obter e estudar dados sobre um sujeito ou um conjunto de sujeitos, a fim de estudar diversos pontos relacionados, conforme o assunto abordado (PRODANOV; FREITAS, 2013).

Quanto aos tipos de procedimentos técnicos, este trabalho foi constituído pelos métodos de pesquisa bibliográfica e estudo de caso, sendo que a pesquisa bibliográfica foi produzida como fundamento teórico do presente trabalho, usando de referências bibliográficas já disponíveis em livros e artigos. O estudo de caso ocorreu descrevendo o projeto Pacto Lajeado pela Paz, e refletindo sobre os dados coletados a partir dele.

4.1 Tecnologias

Para este trabalho foram utilizadas tecnologias buscando alcançar os objetivos definidos para a conclusão com sucesso do mesmo. Desta forma, a seguir serão apresentadas as ferramentas que foram utilizadas.

4.1.1 Ecossistema Python

Python é uma linguagem de programação de alto nível e de fácil compreensão, que incorpora conceitos de orientação a objetos e de programação funcional. Seus ambientes de execução possuem suporte para vários sistemas operacionais, facilitando a adequação aos mais diversos ambientes (DA SILVA et al., 2019).

Segundo Borges (2010), esta linguagem engloba uma ampla coleção de bibliotecas e módulos prontos, permitindo também que *frameworks* externos sejam adicionados. Recursos disponíveis em outras linguagens modernas, como introspecção, metaclasses e testes de unidade, também fazem parte dos recursos de Python.

Embora a linguagem já seja amplamente utilizada em várias áreas, ela também se tornou uma opção muito popular para tarefas de manipulação de dados. Isso ocorreu devido ao aprimoramento de bibliotecas estatísticas, como o Pandas (BORGES, 2010).

Para Mckinney (2012), Pandas é umas das principais bibliotecas do Python, oferecendo estruturas e funções de alto nível para tornar o serviço de manipulação e exploração de dados fácil, rápido e expressivo. Também oferece funcionalidades de alto padrão de indexação, o que facilita remodelar, separar e cortar dados, efetuando seleções e uniões de subconjuntos de dados.

Há dois tipos de estruturas de dados fundamentais no Pandas, segundo Mckinney (2012):

- Series: é semelhante a um vetor de somente uma dimensão, que possui um vetor de dados indexado. A estrutura também pode ser definida como um dicionário de extensão fixa.

- Data Frame: é semelhante a uma planilha eletrônica, possuindo um conjunto de colunas que podem representar diferentes tipos de dados, também indexados.

Outras bibliotecas muito importantes são Matplotlib e Numpy. A biblioteca Matplotlib serve para a visualização de dados, através de gráficos em duas ou três dimensões. Por outro lado, a biblioteca Numpy oferece funções para operações matemáticas entre vetores, álgebra linear e geração de números aleatórios, sendo a base para a construção de funcionalidades mais avançadas (MCKINNEY, 2012).

Já a biblioteca Geopy, auxilia na localização de coordenadas de endereços geográficos, utilizando geocodificadores especializados que acessam serviços externos (GEOPY, 2018).

Estes serviços, como Google Maps Platform, OpenStreetMap Nominatim e Bing Maps, podem ser acessados por APIs específicas. O Geopy é a biblioteca que proporciona o acesso a estes diversos serviços em um pacote único. Desta forma, cada serviço tem seu Termo de Uso, bancos de dados geográficos e preços. O geocodificador Nominatim, por exemplo, é gratuito, porém possui baixos limites de solicitação (GEOPY, 2018).

Segundo Pedregosa et al. (2011), o ecossistema Python ainda conta com o Scikit-learn, que é um módulo que engloba um amplo conjunto de algoritmos de AM para solucionar problemas, permitindo abordagens de aprendizado supervisionado e não-supervisionado. Esse pacote é facilmente integrado com outros pacotes que não fazem parte do escopo de pacotes de análise de dados.

Para uma plotagem interativa dos dados, há a biblioteca Plotly, de código aberto e que possui cerca de 40 variedades de gráficos. Ela possui modelos cobrindo uma vasta diversidade de contextos de uso: financeiros, científicos, estatísticos e tridimensionais (PLOTLY, 2020).

4.1.2 Anaconda

Anaconda é uma distribuição de pacotes científicos de código aberto, focada na linguagem Python. Tipicamente inclui as bibliotecas NumPy, Pandas e também, o ambiente Jupyter para execução local em um microcomputador. Além disso, disponibiliza uma coleção completa de ferramentas de Análise de Dados, além de ser de instalação simples (ANACONDA, 2019).

4.2 Técnicas utilizadas

Foram utilizadas para esse trabalho, tarefas e técnicas de Mineração de Dados, buscando alcançar os objetivos definidos. Desta forma, a seguir serão detalhadas as técnicas empregadas.

4.2.1 Visualização de Dados

Para a apresentação da análise descritiva podem ser utilizadas técnicas de visualização de dados. Segundo Castro e Ferrari (2016), a visualização de dados refere-se à representação visual (gráfica) dos dados, com o propósito de se extrair conhecimento facilmente e rapidamente, e desta forma possibilitar a distribuição do conhecimento com as partes interessadas. As técnicas de visualização auxiliam neste processo de descoberta de conhecimento.

Embora existam várias técnicas de visualização, neste trabalho foram utilizadas apenas as técnicas de histogramas, gráfico de Pareto e gráfico de setores. Castro e Ferrari (2016) explicam que:

- Histograma: é a ilustração da distribuição de frequências dos dados, através de um gráfico de barras de um ou diversos atributos da base. As alturas das barras representam os valores das frequências;

- Gráfico de Pareto: é muito semelhante a um histograma, com barras verticais, porém apresentadas em ordem decrescente de frequência;

- Gráfico de setores: é conhecido também como gráfico do tipo torta, sendo um gráfico circular repartido em setores, com cada setor tendo um tamanho proporcional aos valores exibidos.

4.2.2 K-Means

Este algoritmo pertence à tarefa de análise de grupos. Conforme Varella e Quadrelli (2017), o K-Means recebe como entrada o parâmetro k, referente à quantidade de grupos pretendida. A definição de entrada k, pode ser encontrada por meio do método Elbow, que possui esse nome devido o seu resultado ser semelhante a um braço, onde se busca um cotovelo para determinar um número de grupos k, a serem criados, conforme o número da amostra.

Esse método, começa com a quantidade de *clusters* unitária e busca pelo melhor resultado após cada incremento. Quando não for encontrado mais nenhum benefício no incremento, o melhor resultado de k foi identificado (VARELLA; QUADRELLI, 2017).

De acordo com Castro e Ferrari (2016), após receber como entrada o parâmetro k, referente à quantidade de grupos pretendida, o K-Means divide o conjunto de n objetos em k grupos, para que a correlação dentro de grupo seja alta e a correlação de elementos entre grupos seja baixa.

A correlação dentro do grupo é avaliada levando em consideração o valor médio dos objetos em um conjunto, onde no centro define-se o centróide. Nesta divisão realizada pelo K-Means, cada objeto pertence ao centróide mais próximo (CASTRO; FERRARI, 2016).

Segundo Da Silva et al. (2019), este k centróides são inicialmente determinados aleatoriamente nos grupos. A seguir, são calculadas as distâncias entre os objetos e cada um dos centróides da base, distribuindo-se cada objeto ao centróide mais

aproximado. Na sequência, novos centróides são calculados e atribuídos conforme a média dos objetos do grupo, quando então pode ocorrer um novo posicionamento dos centróides e alocações dos objetos nos grupos. O K-Means termina quando não houver mais reposicionamento de centróides e alocações de objetos.

4.2.3 Apriori

Segundo Faceli et al. (2011), o algoritmo Apriori foi o primeiro para mineração de regras de associação, e engloba gerar os grupos de itens frequentes e extrair regras.

Para Castro e Ferrari (2016), uma maneira de diminuir o consumo computacional dos algoritmos de regras de associação, é separar os requisitos de suporte e confiança mínima das regras. Dado que o suporte da regra necessita somente do grupo de itens e desta forma, grupos de itens poucos frequentes podem ser descartados no início do processo, sem a necessidade de calcular a confiança.

Faceli et al. (2011) destacam que este algoritmo divide os problemas em duas subtarefas:

- Criação do grupo de itens frequentes: encontrar todos os grupos de itens frequentes.

- Criação das regras: utiliza os grupos de itens frequentes para formar regras.

Desta forma, pode se considerar o exemplo que em um grupo ABCD, e AB são frequentes, possibilitando definir que a regra AB → CD pode ser validada calculando a razão, confiança = suporte (ABCD) / suporte (AB). Caso a confiança seja maior ou igual à confiança mínima determinada, então a regra é considerada válida.

Castro e Ferrari (2016) explicam que o primeiro grupo de critérios de análise é determinado com base em premissas estatísticas. Assim, regras que contêm itens respectivamente únicos ou envolvem um número pequeno de operações são irrelevantes. Deste modo, pode-se sugerir de forma objetiva medidas de interesse que classificam tais propriedades das regras, como o suporte e a confiança. O suporte de

uma regra de associação, X → Y, aponta a frequência do evento da regra. A confiança, que também é conhecida como acurácia, investiga a ocorrência da parte decorrente da regra em relação ao precedente, ocasionando o nível de confiança entre os itens.

Demais medidas de interesse envolvem a convicção e o *lift*. Convicção é uma medida da implicação e tem valor um, caso os itens não estejam relacionados. Um valor alto de convicção, significa que o consequente depende muito do antecedente. O *lift* é a probabilidade de Y ocorrer quando X está presente, assim quando o *lift* for igual a um, X não causa impacto em Y, e se o *lift* for maior que um, existe uma relação entre X e Y (FACELI et al., 2011).

Sharma e Tiwari (2014), ainda comentam sobre o *leverage*, que é a medida que apresenta a diferença entre a frequência observada de X e Y aparecendo juntos e a frequência que seria esperada, se X e Y forem independentes. Valores iguais ou abaixo do valor zero, indicam uma forte independência entre X e Y. Quando próximos a um, indicam uma forte regra de associação.

4.3 Desenvolvimento

Para o presente trabalho, foi implementado um processo de Mineração de Dados seguindo a metodologia CRISP-DM, utilizando como estudo de caso um conjunto de dados fornecido pelo projeto Pacto Lajeado pela Paz.

O Pacto Lajeado pela Paz é uma iniciativa da Prefeitura Municipal de Lajeado - RS com demais entidades, lançado oficialmente em junho de 2019. O Pacto é uma ação que envolve diversos setores, e tem como propósito incentivar e promover a cultura de paz na cidade. Ele ainda conta com uma empresa terceirizada que atua na área de segurança pública. O objetivo central é desenvolver um sistema municipal para a prevenção de violência em Lajeado (LAJEADO, 2019).

A partir de contatos feitos junto aos provedores do projeto, identificou-se uma prioridade: a necessidade de sistematizar, agregar e tratar dados importantes. Visto que o projeto estava em processo inicial e que estão envolvidas várias pastas municipais, como a Secretaria da Saúde, Secretaria da Educação, Secretaria do

Esporte e Cultura, Secretaria de Trabalho, Habitação e Assistência Social (STHAS) e Secretaria Municipal de Segurança Pública, foi levantado que vários indicadores necessários dependiam de dados do município, mantidos de forma esparsa.

Observou-se que cada pasta possui seu próprio repositório de dados, sendo que algumas possuem sistemas específicos para gerenciá-los, enquanto outras necessitam digitar seus dados em planilhas para poderem fazer análises próprias. Ainda, em algumas situações, os dados relevantes são mantidos em papel.

Com bases tão distintas e de diferentes origens, este trabalho pretendeu propor uma estrutura simples de coleta, organização e manipulação de alguns dados específicos, para que análises mais complexas pudessem ser realizadas. Para isso, a forma que este trabalho foi desenvolvido está detalhada nas seções a seguir

4.3.1 Entendimento dos Negócios

Esta etapa foi realizada através de reuniões regulares com as pastas vinculadas à Prefeitura de Lajeado - RS, que atuam de forma integrada no âmbito do projeto Pacto Lajeado pela Paz. Destas reuniões participaram os coordenadores do projeto e gestores das secretarias vinculadas à Prefeitura, incluindo os técnicos responsáveis pelos abastecimentos dos dados.

Desta forma, foi analisado em conjunto o cenário do projeto Pacto Lajeado pela Paz, registrando as expectativas de cada secretaria e da organização. Em especial, foram mapeados indicadores estatísticos com embasamento científico, que deveriam ser construídos prioritariamente, com o intuito de dar base para ações posteriores de repressão ou prevenção na comunidade.

4.3.2 Entendimento dos Dados

Uma planilha *online* foi montada pela Prefeitura de Lajeado e compartilhada com a Secretaria da Saúde, Secretaria da Educação, STHAS e Polícia Civil com o intuito de que as próprias instituições informassem os seus dados. No entanto,

analisando a planilha, percebeu-se que os dados estavam sumarizados, havendo pouca informação detalhada. Optou-se, então, por extrair e analisar os dados brutos, obtidos diretamente das fontes, ou seja, de cada secretaria. Um dos objetivos foi visitar e identificar os meios em que esses dados estão armazenados.

Com as visitas às secretarias, ficou clara a forma em que cada pasta armazena seus dados.

Na STHAS, os dados eram recebidos no formato de papel e ainda não possuíam um sistema para inseri-los. Então, o próprio Assistente Social montou uma base em formato de planilha eletrônica para que pudesse fazer suas análises. Contudo, lamentou as poucas informações que eram encaminhadas pelas entidades registradoras dos casos.

As únicas informações que a STHAS possui, utilizadas pelo Pacto Lajeado pela Paz, são dados categorizados sobre jovens com medidas socioeducativas, apresentando o mês (em que a medida foi iniciada), qual medida, idade do jovem e escola (se estiver em uma escola).

Na Secretaria da Educação, as escolas municipais possuem um aplicativo, chamado Nota 10. Os dados são inseridos pelos próprios professores e o armazenamento destes dados fica no *data center* do setor de Tecnologia da Informação, da Prefeitura de Lajeado. O acesso aos dados foi permitido pela Prefeitura e coordenadores do Projeto, porém o setor de Tecnologia da Informação não disponibilizou a base desses dados.

Já a Secretaria da Saúde, recebe os dados em formato de papel, em formulários, e insere estes dados no sistema, chamado Sistema de Informação de Agravos de Notificação (SINAN). Os formulários recebidos contam com informações de violências e agressões sofridas por Lajeadenses, tanto na cidade de Lajeado, quanto em outras cidades. Quem os envia são unidades de saúde ou assistência social. Esta base de dados pôde ser facilmente extraída através do próprio sistema, em um arquivo com o formato Data Base File (DBF).

Para compreender cada atributo da base extraída da Secretaria da Saúde, assim como compreender os seus objetos, foi mantido contato com os técnicos

responsáveis por alimentar os dados. Também o próprio formulário SINAN (ANEXO A) utilizado para registrar as violências, se mostrou muito útil para o entendimento da base.

4.3.3 Preparação dos Dados

Inicialmente, a base de dados extraída da Secretaria da Saúde em formato DBF do sistema SINAN foi convertida para formato CSV. A base trabalhada possui registros do ano de 2010 até 2019, com aproximadamente 4.000 linhas e 198 colunas, contendo informações gerais da pessoa vitimada (não contendo o nome), tipo de violência e vínculo com o possível agressor, unidade registradora, unidade a ser encaminhada, além de demais dados gerais, como por exemplo, data da digitação.

Os dados extraídos foram padronizados para que fosse possível realizar a execução das técnicas de Mineração de Dados. Usando um *notebook*, criado no *framework* Jupyter e executado na plataforma Anaconda, foram inicialmente importadas as bibliotecas necessárias. A seguir a base de dados da Secretaria da Saúde foi importada, sendo que o atributo “NU_NOTIFIC,C,7”, referente ao número da notificação, foi definido como índice para fins de rastreabilidade (ou seja, caso seja necessário olhar casos em particular), como mostra a Figura 5.

Figura 5 – Importação da base de dados da Secretaria da Saúde

```
saude = pd.read_csv('VIOLENET.csv', sep=';', dtype=tipos)
saude.set_index('NU_NOTIFIC,C,7')
```

Fonte: Da autora (2020).

Cada atributo foi individualmente analisado e os desnecessários excluídos, conforme pode ser visto na Figura 6. Os atributos definidos como desnecessários foram: atributos com informações particulares da vítima, como número de telefone e nome da mãe, que foram excluídos para anonimiza-las; atributos com informações administrativas da entidade que registrou a violência; atributos de campo texto aberto com observações; atributos que apresentavam o registro do procedimento feito na vítima após a violência, como nos casos de violência sexual; e unidade de encaminhamento da vítima. No total 157 atributos foram excluídos, permanecendo

apenas 41. A maior parte dos atributos descartados eram categóricos, representando apenas alguns valores discretos de listas predefinidas

Figura 6 – Exclusão dos atributos desnecessários

```
saude = saude.drop(columns=['TP_NOT,C,1','ID_AGRAVO,C,4','ID_UNIDADE,C,7','DT_NOTIFIC,D','SEM_NOT,C,6','SG_UF_NOT,C,2','ID_MUNIC
                            'NM_PACIENT,C,70','CS_RACA,C,1','ID_CNS_SUS,C,15','NM_MAE_PAC,C,60','SG_UF,C,2','ID_MN_RESI,C,6','ID_D
                            'DS_REF_RES,C,70','NU_CEP,C,8','DDD,C,2','FONE,C,9','ZONA,C,1','ID_PAIS,C,4','DT_INVEST,D','DEF_FISICA
                            'SG_UF_OCOR,C,2','ID_MN_OCOR,C,6','ID_DIS_OCO,N,8,0','ID_LOG_OCO,N,8,0','DS_COMP_OC,C,15','DS_REF_OCO,
                            'SEX_PUDOR,C,1','SEX_PORNO,C,1','SEX_EXPLO,C,1','SEX_OUTRO,C,1','SEX_ESPEC,C,30','PEN_ORAL,C,1','PEN_A
                            'PROC_SEMEN,C,1','PROC_VAGIN,C,1','PROC_CONTR,C,1','PROC_ABORT,C,1','NUM_ENVOLV,C,1','REL_SEXUAL,C,1',
                            'REL_CONHEC,C,1','REL_CUIDA,C,1','REL_PATRAO,C,1','REL_INST,C,1','REL_POL,C,1','REL_PROPRI,C,1','REL_O
                            'MPU,C,1','DELEG_CRIA,C,1','DELEG_MULH,C,1','DELEG,C,1','INFAN_JUV,C,1','DEFEN_PUBL,C,1','DT_ENCERRA,D
                            'LESAO_NAT,C,2','LESAO_ESPE,C,30','LESAO_CORP,C,2','REDE_SAU,C,1','ASSIST_SOC,C,1','REDE_EDUCA,C,1','A
                            'DIR_HUMAN,C,1','MPU,C,1','DELEG_CRIA,C,1','DELEG_MULH,C,1','DELEG,C,1','INFAN_JUV,C,1','DEFEN_PUBL,C,
                            'ENC_DPCA,C,1','ENC_DELEG,C,1','ENC_MPU,C,1','ENC_MULHER,C,1','ENC_CREAS,C,1','ENC_IML,C,1','ENC_OUTR,
                            'DT_TRANSSM,D','DT_TRANSRM,D','DT_TRANSRS,D','DT_TRANSSE,D','NU_LOTE_V,C,7','NU_LOTE_H,C,7','IDENT_MIC
```

Fonte: Da autora (2020).

Para analisar a qualidade dos dados, foi verificado os números de dados ausentes em cada atributo. A Figura 7, mostra metade dos atributos da base final, a título de ilustração.

Figura 7 – Análise de colunas com valores ausentes

```
saude.isnull().sum()

NU_NOTIFIC,C,7          0
NU_ANO,C,4              0
DT_OCOR,D               1
DT_NASC,D              23
NU_IDADE_N,N,4,0        0
CS_SEXO,C,1             0
CS_GESTANT,C,1          0
CS_ESCOL_N,C,2         68
NM_BAIRRO,C,60         64
ID_OCUPA_N,C,6       2422
SIT_CONJUG,C,1         23
DEF_TRANS,C,1           5
NM_BA_OCOR,C,30       901
NO_LOG_OCO,C,60       863
NM_LOG_RES,C,7       1493
HORA_OCOR,C,5        3157
LOCAL_OCOR,C,2          7
OUT_VEZES,C,1           3
LES_AUTOP,C,1           6
VIOL_FISIC,C,1         10
VIOL_PSICO,C,1         37
```

Fonte: Da autora (2020).

É possível averiguar que grande parte da base teve que ser individualmente analisada e tratada, para se tornar coerente. Para isso foi criada uma função genérica, chamada *coding*, que utiliza a função de substituição, conforme mostra a Figura 8.

Figura 8 – Função para substituição de valores

```
def coding(col, codeDict):
  colCoded = pd.Series(col, copy=True)
  for key, value in codeDict.items():
    colCoded.replace(key, value, inplace=True)
  return colCoded
```

Fonte: Da autora (2020).

Essa função foi utilizada para tratar atributos nominais como, por exemplo, o nome dos bairros, que estavam com nomes escritos incorretamente ou com variações de nomes. Por exemplo, a base de dados continha os nomes: MOINHOS DA ÁGUA, MUINHOS DA AGUA, MOINHOS DAGUA, MOINHOS DE ÁGUA, MOINHOS D' ÁGUA e MOI%, estes foram corrigidos para: MOINHOS DA AGUA. Todos os nomes de bairros foram tratados para que no fim ficassem somente os 27 bairros da cidade de Lajeado, informados corretamente, como mostra a Figura 9.

Figura 9 – Tratamento do nome dos bairros

```
saude['NM_BA_OCOR,C,30'] = coding(saude['NM_BA_OCOR,C,30'], {'CA%':'CAMPESTRE','CAM%':'CAMPESTRE','CAR%':'CARNEIROS','A%':'Ameri
                                                            'CANEIROS':'CARNEIROS','CEN%':'CENTENARIO','CENTENÁRIO':'CENTENARIO
                                                            'CE%':'CENTENARIO','CENTENARO':'CENTENARIO','CENTOR':'CENTRO','CON%
                                                            'FLORE%':'FLORESTAL','HI%':'HIDRAULICA','HIDRÁULICA':'HIDRAULICA','
                                                            'JARDIM DO CREDRO':'JARDIM DO CEDRO','JARD%':'JARDIM DO CEDRO','JA%
                                                            'JARDIN DO CEDRO':'JARDIM DO CEDRO','MOINHOS DAGUA':'MOINHOS DA AGU
                                                            'MOINHOS D´ ÁGUA':'MOINHOS DA AGUA','MOINHOS DE AGUA':'MOINHOS DA A
                                                            'MOI%':'MOINHOS DA AGUA','MON%':'MONTANHA','MOR%':'MORRO 25','MOINH
                                                            'SANTA ANTONIO':'SANTO ANTONIO','SAMTO ANTONIO':'SANTO ANTONIO','SA
                                                            'SAO CRISTÓVÃO':'SAO CRISTOVAO','SAO CRITOVAO':'SAO CRISTOVAO','DOI
                                                            'MAONTANHA':'MONTANHA','UNIVERSITARO':'UNIVERSITARIO','JARDIM DE CE
                                                            'JARDIM CEDRO':'JARDIM DO CEDRO','IMIGRANTES':'IMIGRANTE','INDUATRI
```

Fonte: Da autora (2020).

Os registros que não possuíam informações do nome do bairro, tiveram a informação 'NAO INFORMADO' inserida. Embora esta informação fosse irrelevante e ignorada no momento da análise dos dados, o registro/objeto precisava ser mantido, devido às demais informações que possuía em outros atributos.

Além da correção dos atributos nominais, devido à grande quantidade de atributos numéricos com dados ausentes, estes também foram tratados. Os atributos numéricos que possuíam valores ausentes, tiveram um 'zero' inserido e os atributos que possuíam demais valores, também foram analisados. Para valores considerados desnecessários, também houve a substituição por 'zero'.

Por exemplo, atributos que possuíam a informação de 'um' para sim, 'dois' para não e 'nove' para ignorada, tiveram seus dados como 'dois' e 'nove' substituídos por

‘zero’, visto que não fazia sentido, para a análise, mantê-los separados. O resultado, portanto, foi apenas valores ‘zero’ e ‘um’, como pode ser visto na Figura 10.

Figura 10 – Tratamento dos dados ausentes e desnecessários

```
saude.update(saude['VIOL_FISIC,C,1'].fillna(0))
saude['VIOL_FISIC,C,1'] = coding(saude['VIOL_FISIC,C,1'], {2:0})
saude['VIOL_FISIC,C,1'] = coding(saude['VIOL_FISIC,C,1'], {9:0})
```

Fonte: Da autora (2020).

Por último, visto que na base de dados havia a informação da idade das vítimas, optou-se por criar no Data Frame um atributo chamado faixa_etaria com o intervalo de 10 anos. Por meio disso seria mais fácil visualizar a distribuição do perfil das vítimas, conforme a idade.

Assim, após cada atributo ser individualmente analisado, novamente foi executado o código para contagem de dados ausentes. Com todos os registros completos, a base final foi exportada como um segundo arquivo, conforme mostra Figura 11.

Figura 11 – Exportando a base final em um segundo arquivo

```
saude.to_csv('base_saude.csv')
```

Fonte: Da autora (2020).

Esta base final foi chamada de ‘base_saude’, sendo a base de dados utilizada para toda a Análise e Mineração de Dados realizada. Após a escolha dos algoritmos e aplicação das demais etapas do CRISP-DM, sempre que era identificado que um dado precisava ser novamente tratado, voltava-se para esse *notebook*. Então tratavam-se os dados ou criava-se um atributo, conforme a necessidade, gerando-se novamente uma segunda base de dados atualizada.

4.3.4 Modelagem

Na etapa de modelagem muitos algoritmos foram analisados e executados, porém nem todos tiveram conhecimento útil extraído ou sucesso em sua execução.

As tarefas de Mineração de Dados utilizadas foram: análise descritiva de dados, análise de grupos e associação.

A primeira análise feita sobre os dados, foi uma análise estatística utilizando a tarefa de MD, chamada análise descritiva dos dados. Um novo *notebook* no *framework* Jupyter e executado na plataforma Anaconda foi criado, para que as principais características dos dados pudessem ser exploradas, simplificadas e descritas.

Posteriormente foi desenvolvido um mapa de calor. Na busca por ruas com maior densidade de violência registradas na cidade de Lajeado, foram utilizadas informações geolocalizadas. Novamente foi criado um *notebook*, em que as coordenadas das ruas foram descobertas e o mapa de calor foi implementado.

A segunda tarefa de MD realizada foi a análise de grupos. Para a realização desta tarefa foi utilizado o algoritmo K-Means. Um novo *notebook* foi criado para que se pudessem identificar características ou combinações de características similares em alguns objetos.

A terceira tarefa de Mineração de Dados realizada, foi a associação. Para a realização desta tarefa foi utilizado o algoritmo Apriori. Um novo *notebook* foi desenvolvido com o objetivo de buscar relações entre os atributos dos objetos.

4.3.5 Avaliação

Nesta etapa, os resultados obtidos foram extraídos de um *notebook* em formato de Portable Document Format (PDF), e enviados por *e-mail* para os especialistas da área da saúde. O objetivo era de que os especialistas validassem, se os resultados faziam sentido e se eram suficientes para uma tomada de decisão.

No entanto, no ano de 2020 sofremos com a pandemia da Covid-19. Como os especialistas envolvidos, atuavam diretamente na área epidemiológica, esta análise especializada não foi possível, devido ao alto volume de demanda para a área da Saúde naquele momento.

4.3.6 Implantação

Foram avaliadas algumas bibliotecas de Python para esta etapa, optando-se pela utilização da biblioteca Plotly, para auxiliar na apresentação interativa dos resultados para os gestores. Essa biblioteca foi escolhida, dada a variedade de interações que ela oferece, como o *download* da imagem, recurso de *zoom* e auto escala, gerados automaticamente para cada gráfico, além de outras maneiras de seleção de dados sobre o gráfico.

Para facilitar a exploração e entendimento dos gráficos, utilizou-se a biblioteca para criar figuras interativas que permitem usar o *mouse* sobre os dados, indicando os exatos valores representados. Também é possível dar *zoom* na imagem e ver os objetos mais próximos ou mais distantes, além de ser possível selecionar um ponto específico para enquadrar no gráfico.

5 TESTES E ANÁLISE DOS RESULTADOS

Neste capítulo são apresentados os resultados obtidos na execução dos procedimentos, apresentados na Seção 4.3, que buscaram através da Mineração de Dados realizar análises e extração de conhecimento dos dados da Secretaria da Saúde. Para melhor organização e entendimento dos resultados, estes foram divididos em subseções.

5.1 Análise descritiva de dados

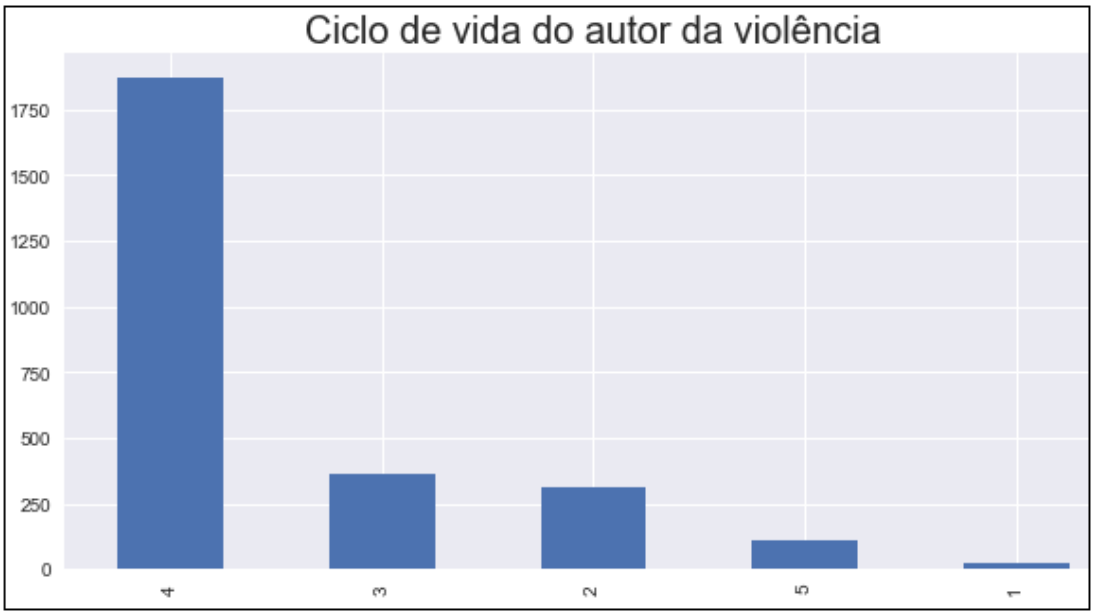
Inicialmente, as bibliotecas relacionadas e a base final com os dados tratados foram importadas em um *notebook*.

A primeira abordagem feita foi uma análise do ciclo de vida do autor da violência. Não há muitas informações sobre os autores, apenas os relatos em que a própria vítima informa, visto que o autor em muitos casos é um desconhecido da vítima. Desta forma, foi criado e executado um código que ignora os objetos/registros, com o valor ‘zero’ e cria um gráfico de Pareto, onde são apresentadas em ordem decrescente, as frequências com que cada ciclo de vida aparece na base.

Neste gráfico de Pareto podemos levar em consideração a categorização do ciclo de vida estabelecida no formulário SINAN, onde o valor ‘um’ representa autores considerados crianças (0 a 9 anos), ‘dois’ representa os adolescentes (10 a 19 anos), ‘três’ representa os jovens (20 a 24 anos), ‘quatro’ representa os adultos (25 a 59

anos) e ‘cinco’ representam os idosos (60 anos ou mais). No gráfico, exibido na Figura 12, é possível visualizar que o ciclo de vida que mais praticou atos de violência é de pessoas adultas, entre 25 a 59 anos, pois aparece com mais frequência na base de dados.

Figura 12 - Gráfico do ciclo de vida do autor da violência

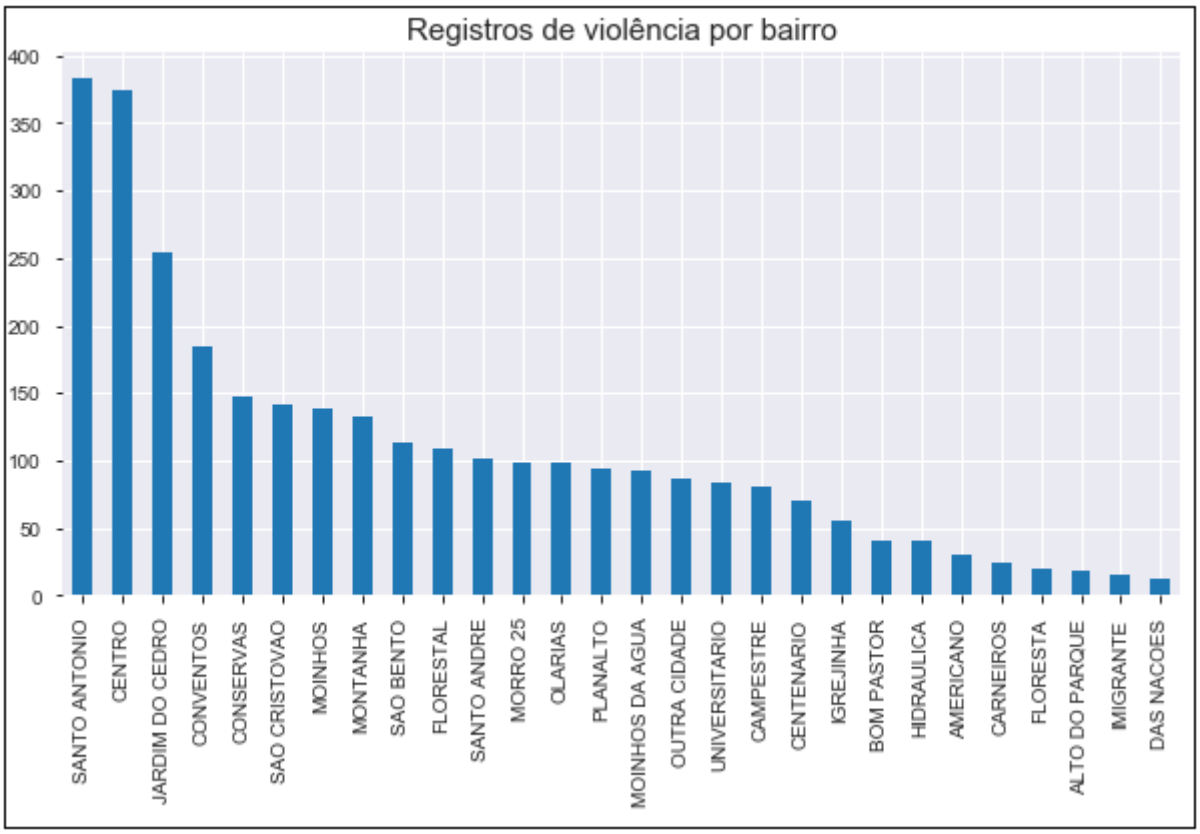


Fonte: Da autora (2020).

Outra análise feita, foi do índice de notificação de violência, agrupado por bairros. Para isso, foi criado e executado um código para agrupar os registros por cada bairro, que ignora os registros com o nome “NAO INFORMADO”, ao invés do nome do bairro, visto a irrelevância desta informação. Após, foi criado um gráfico de Pareto que apresenta em ordem decrescente os bairros, conforme a quantidade de registros de violência nestes ocorridos.

Na Figura 13 é possível verificar quais os bairros que possuem mais ações violentas registradas. Os três bairros com maior número de registros são: em primeiro lugar o bairro Santo Antônio, em segundo o bairro Centro e na sequência o bairro Jardim do Cedro.

Figura 13 - Gráfico do número de notificações por bairro



Fonte: Da autora (2020).

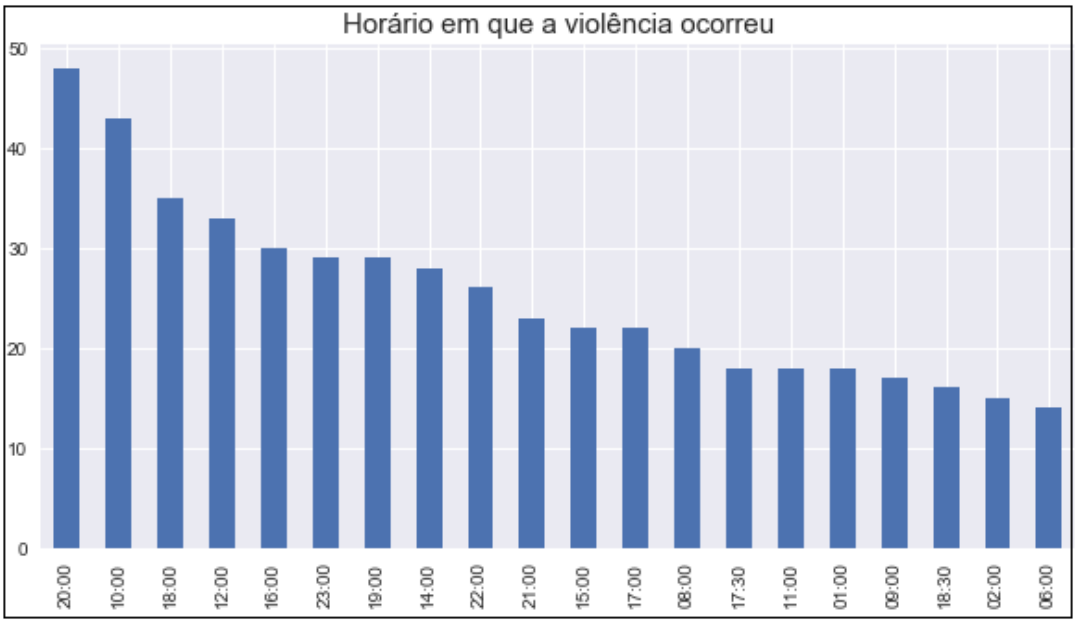
Na Figura 13, também se encontra a opção ‘OUTRA CIDADE’. De acordo com a etapa de entendimento dos dados feita com a Secretaria da Saúde, esta base possui registros de Lajeadenses, que sofreram violência fora de Lajeado e que também foram contabilizados. Mesmo que não haja ação sobre a violência em outras cidades, a informação foi mantida para fins de conhecimento da área.

Embora não houvesse muitos registros contendo o horário em que a violência ocorreu, também foi gerado um gráfico de Pareto em ordem decrescente, para apresentar uma ideia preliminar dos horários em que mais ocorreram violências. Para isso, foi criado um código que faz o agrupamento por horário (arredondado para a hora anterior) e ignorando os valores com a opção de hora ‘zero’, visto que se trata de valores anteriormente nulos. Assim, permanecendo a hora ‘00:00’, que é um horário válido.

Na Figura 14 é possível verificar que o horário que aparece com mais frequência é das 20 horas. Em uma visão geral, se compararmos os horários que

aparecem listados no *ranking* dos horários com mais frequência de registros, o turno em que mais ocorreu violências é a noite, entre 18:00 horas e meia-noite.

Figura 14 - Gráfico do agrupamento dos registros por horário



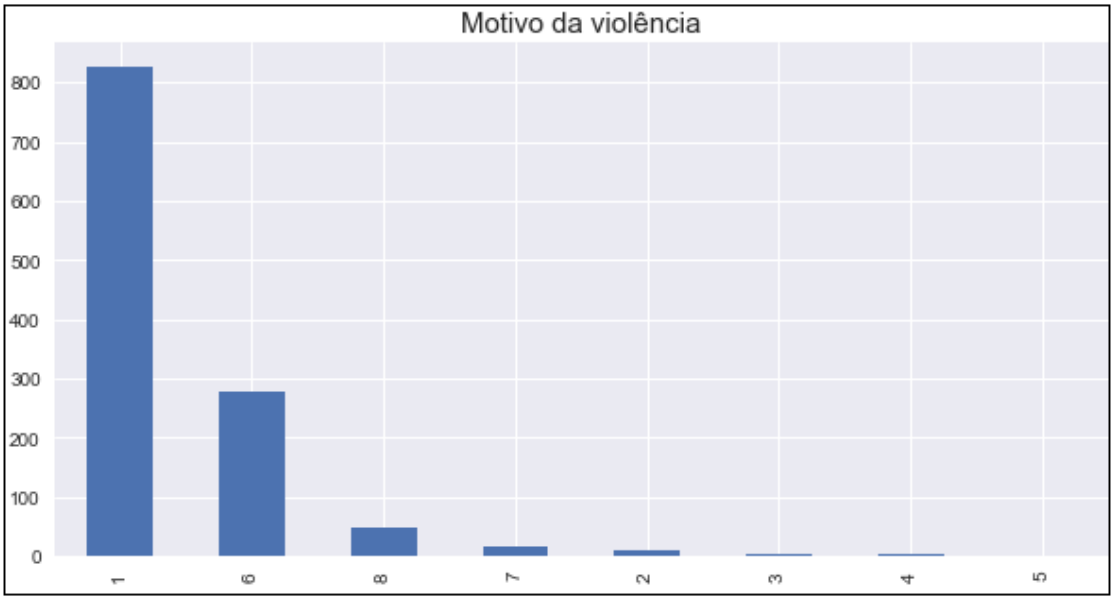
Fonte: Da autora (2020).

Outra análise interessante realizada, foi o agrupamento de registros por motivos da violência. Para este gráfico foi criado um código que faz o agrupamento e que também ignora o valor ‘zero’, gerando um gráfico de Pareto em ordem decrescente, conforme a quantidade de registros que cada motivo de violência possui.

Nesta análise, é levado em consideração que o valor ‘um’ representa o motivo de violência chamado sexismo, ‘dois’ representa homofobia/lesbofobia, ‘três’ representa racismo, ‘quatro’ representa intolerância religiosa, ‘cinco’ representa xenofobia, ‘seis’ representa conflito geracional, ‘sete’ representa situação de rua e ‘oito’ representa deficiência como motivo da violência.

Nota-se, que o motivo de violência ‘sexismo’ aparece com maior frequência na base de dados, possuindo mais de 800 registros de violências relatadas (em um universo de mais de 4.000 casos). Em segundo, temos o motivo ‘conflito geracional’ com quase 300 registros. E por terceiro, temos o motivo da violência ocasionada, devido alguma ‘deficiência’ da vítima, com aproximadamente 50 registros, conforme ilustra a Figura 15.

Figura 15 - Agrupamento dos registros por motivo

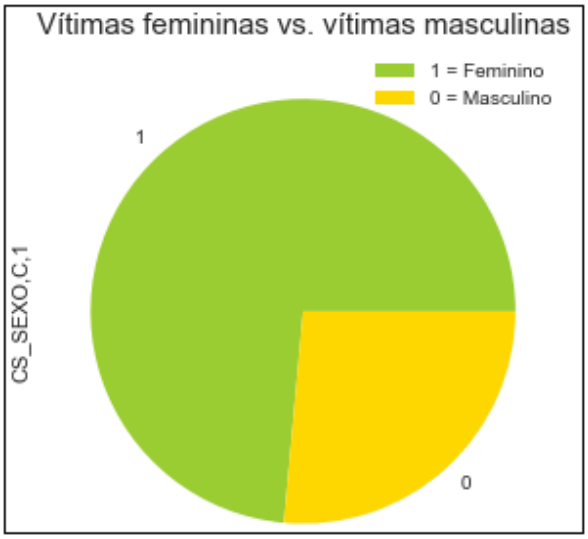


Fonte: Da autora (2020).

Com base na análise anterior, em que o maior motivo de violência registrado em Lajeado é o sexismo, identificou-se ser relevante, a análise destes registros por sexo das vítimas. Desta forma, foi criado e executado um código para geração de um gráfico de setores, onde é apresentado visualmente com setores em amarelo, a quantidade de registros de vítimas masculinas e em verde, as vítimas femininas.

No gráfico de setores da Figura 16, verifica-se que aproximadamente 75% das vítimas são femininas e aproximadamente 25% das vítimas são masculinas.

Figura 16 - Agrupamento dos registros por sexo das vítimas



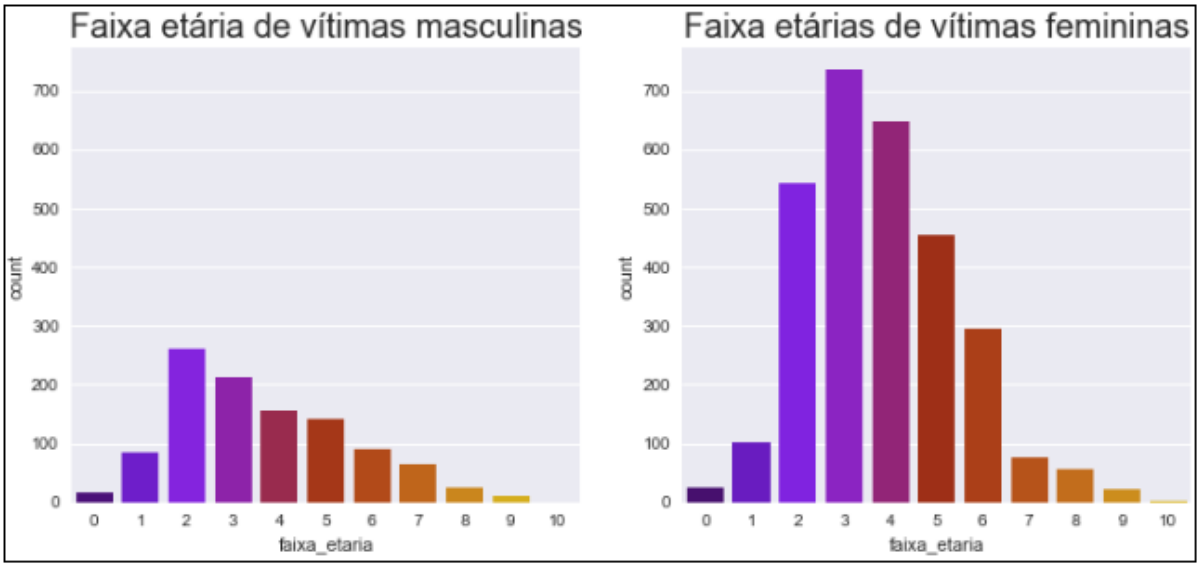
Fonte: Da autora (2020).

Com a descoberta de conhecimento do maior motivo de violência e do grande número de vítimas femininas, passou-se a analisar os dados por sexo separadamente. Desta forma, para analisar a faixa etária destas vítimas, foram criados dois histogramas, um para visualização do agrupamento das faixas etárias das vítimas masculinas e outro para visualização dos agrupamentos das faixas etárias das vítimas femininas.

Nos histogramas apresentados na Figura 17, considera-se para ambos os casos, o valor ‘zero’ para vítimas com menos de um ano de idade, ‘um’ para vítimas de um a nove anos de idade, ‘dois’ para vítimas com 10 a 19 anos de idade, ‘três’ para vítimas com 20 a 29 anos de idade, ‘quatro’ para vítimas com 30 a 39 anos de idade, ‘cinco’ para vítimas com 40 a 49 anos de idade, ‘seis’ para vítimas com 50 a 59 anos de idade, ‘sete’ para vítimas com 60 a 69 anos de idade, ‘oito’ para vítimas com 70 a 79 anos de idade, ‘nove’ para vítimas com 80 a 89 anos de idade, ‘dez’ para vítimas com 90 a 99 anos de idade e ‘onze’ para vítimas com mais de 100 anos (este último não aparece no gráfico, pois não há casos envolvendo essa faixa etária).

Na Figura 17, de acordo com o primeiro gráfico pode-se dizer que a faixa etária das vítimas masculinas, com mais registros é a de 10 a 19 anos de idade. Já no segundo gráfico, podemos considerar que a faixa etária das vítimas femininas com mais registros é a de 20 a 29 anos.

Figura 17 - Gráfico do agrupamento dos registros por faixa etária das vítimas

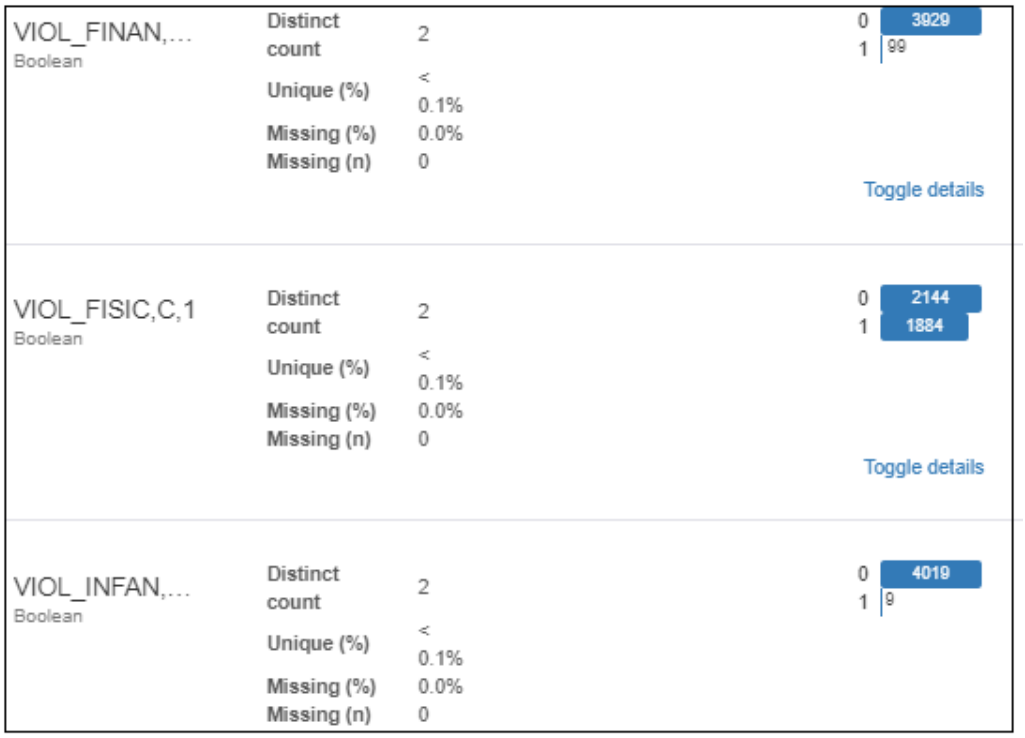


Fonte: Da autora (2020).

Para identificar a violência que mais ocorreu, foi utilizada a biblioteca Pandas Profiling, que é uma extensão da biblioteca Pandas. Executado esta biblioteca, foi gerado um relatório de visão global sobre dados analisados, sendo possível a impressão deste relatório.

Dentro do relatório gerado, é possível visualizar para cada atributo um histograma. Através desta informação, verificou-se que a violência que mais tem registros é a violência física, com 1884 registros. Para o entendimento destes dados, leva-se em consideração de que ‘zero’ é não (determinada violência não ocorreu) e ‘um’ é sim (determinada violência ocorreu). Na Figura 18, são apresentados somente três tipos de violência, porém também existem outras, como a violência psicológica, ou negligência, além de algumas agressões, como verbal, uso de força ou envenenamento.

Figura 18 - Histograma dos registros de violência



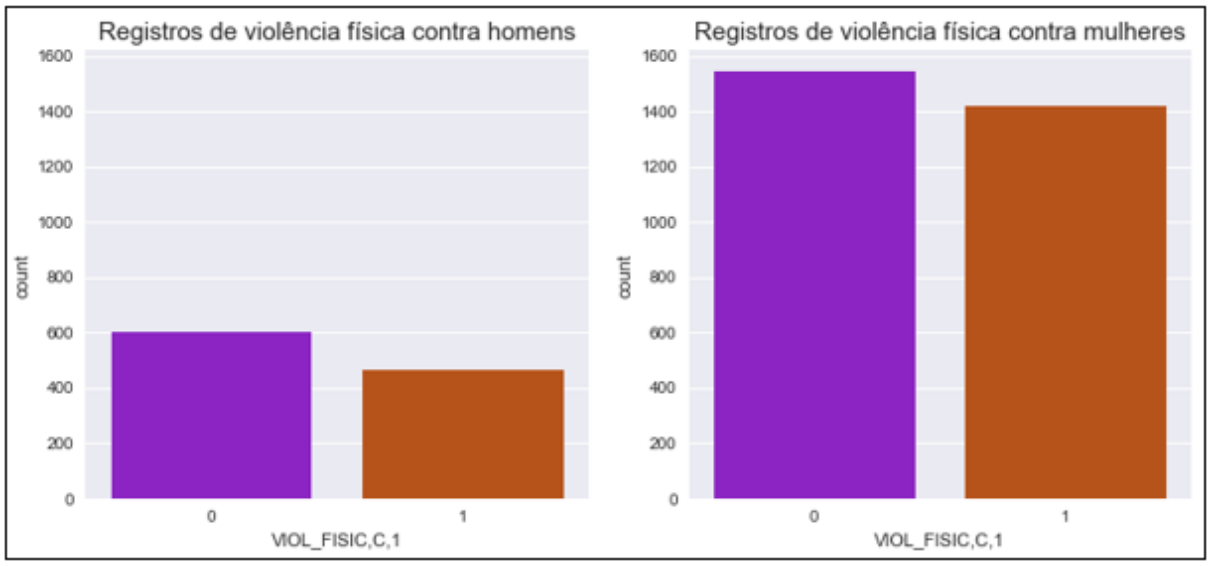
Fonte: Da autora (2020).

Para aprofundar ainda mais o conhecimento, visto que o sexo feminino possuía maior número de registros de violência e que a violência mais registrada foi a física, além do principal motivo da violência ter sido o sexismo, foram criados dois novos

histogramas, para que se pudesse analisar o número de registros de violência física por sexo da vítima.

Na Figura 19, o primeiro gráfico representa os registros de violência física contra homens, onde ‘zero’ são os registros, onde não houve violência física (ocorreu outro tipo de violência), e ‘um’ houve violência física contra homens. O segundo gráfico, representa os registros de violência física contra mulheres, onde ‘zero’ são os registros em que não houve violência física (aconteceu outro tipo de violência), e ‘um’ houve violência física contra mulheres. Com estes dois gráficos é possível visualizar que há um grande volume de registros de violência física contra homens, mas que o maior volume é contra mulheres.

Figura 19 - Gráficos agrupando os registros de violência por sexo



Fonte: Da autora (2020).

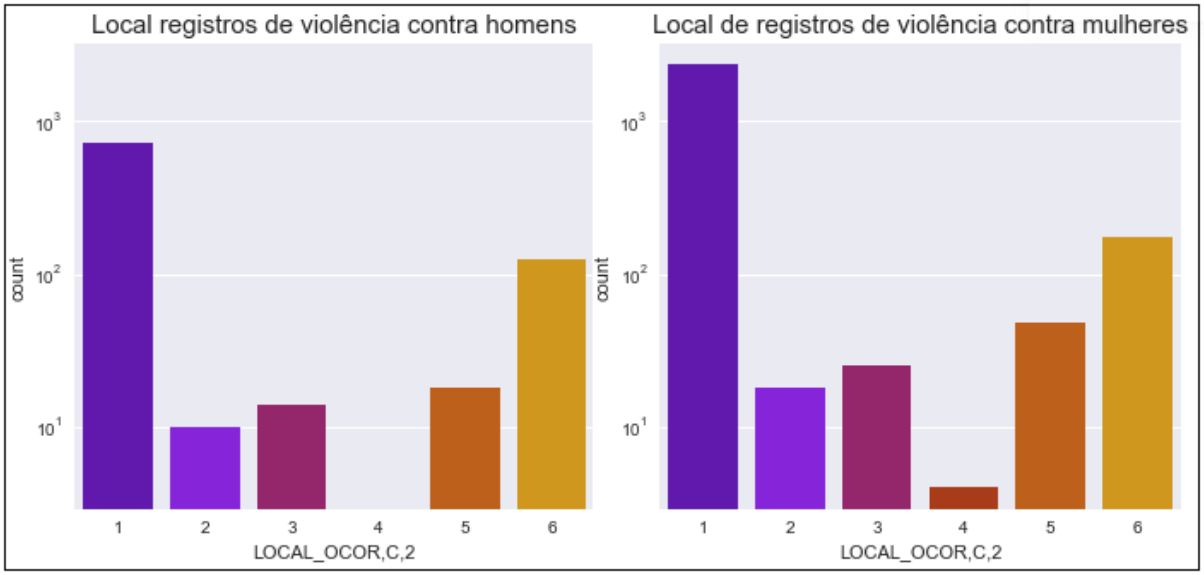
Como a base possuía a informação do local onde a violência ocorreu, foi executado um código que cria também dois histogramas, sobre as vítimas masculinas e sobre as vítimas femininas. O objetivo foi apresentar o local em que mais ocorreram violências.

Na Figura 20, o primeiro gráfico apresenta o agrupamento dos registros de violência contra homens por local do ocorrido, enquanto o segundo é referente às mulheres. Para ambos os gráficos, considera-se ‘um’ como local da violência sendo a própria residência da vítima, ‘dois’ como local da violência uma habitação coletiva,

‘três’ sendo o local a própria escola da vítima, ‘quatro’ sendo um local de prática esportiva, ‘cinco’ sendo um bar ou similar e ‘seis’ sendo uma via pública.

Devido à desproporção dos valores, havia um valor muito mais alto de registros no local da própria residência da vítima, para ambos os sexos e por isso pode-se constatar na Figura 20, que os gráficos utilizam escala logarítmica, para facilitar a visualização dos locais com menores números de registros.

Figura 20 - Gráficos agrupando os registros pelo local em que a violência ocorreu



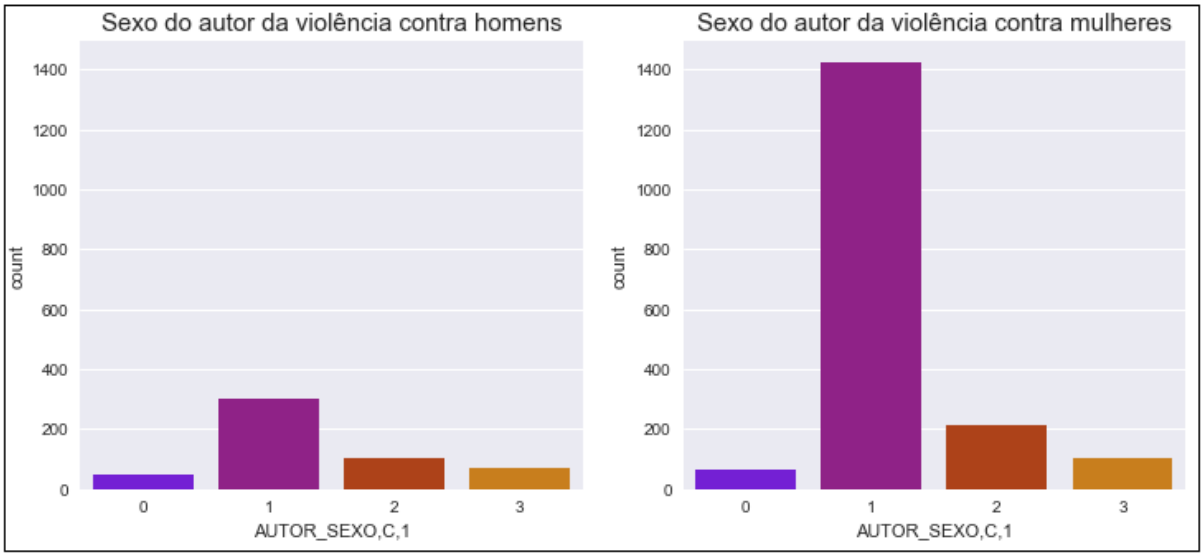
Fonte: Da autora (2020).

Por último, foram novamente gerados dois histogramas, sobre o sexo masculino e sobre o sexo feminino, onde o agrupamento feito foi referente ao sexo do autor da violência, podendo ser masculino, feminino ou ambos os sexos.

Assim, a Figura 21 apresenta dois gráficos, montados com o objetivo de entender as implicações do sexo do agressor. No primeiro gráfico, considera-se ‘um’ os casos em que um homem cometeu violência contra outro homem, considera-se ‘dois’ os casos em que uma mulher cometeu violência contra um homem e ‘três’ para os casos em que ambos os sexos cometeram violência contra um homem. Já no segundo gráfico, considera-se ‘um’ para os casos em que um homem cometeu violência contra uma mulher, ‘dois’ para os casos em que uma mulher cometeu violência contra outra mulher e ‘três’ para os casos em que ambos os sexos cometeram violência contra uma mulher. Em ambos os gráficos, é possível visualizar

que o sexo do autor que mais tem registros como autor da violência é o masculino, em especial em relação a vítimas femininas.

Figura 21 - Gráficos agrupando os registros pelo sexo do autor da violência



Fonte: Da autora (2020).

5.2 Georreferenciamento da localização dos casos

Para que o georreferenciamento ocorresse, foi implementado um novo *notebook* criado no *framework* Jupyter, importando a base corrigida dos dados e processando o Data Frame criado com as ruas das violências registradas.

Através da biblioteca Geopy foi utilizada o geocoder Nominatim, para que fossem obtidas a latitude e a longitude das ruas de Lajeado para cada registro da base de dados. A seguir, o código criou um arquivo com três colunas, o nome da rua, a longitude e a latitude destas ruas para utilizações futuras destas informações, como pode ser visto na Figura 22.

Figura 22 - Código para a obtenção da geolocalização

```
locations=pd.DataFrame({"Name":saude['NO_LOG_OCO,C,60']})
lat_lon=[]
cont=0
geolocator=Nominatim(user_agent="app", timeout=3)
for location in locations['Name']:
    d = {'street':location , 'city':'Lajeado', 'state':'Rio Grande do Sul', 'country':'Brasil'}
    location = geolocator.geocode(d)
    if location is None:
        lat_lon.append(np.nan)
    else:
        geo=(location.latitude,location.longitude)
        lat_lon.append(geo)
        cont+=1
locations['Latitude', 'Longitude']=lat_lon
locations.to_csv('locations.csv',index=False)
```

Fonte: Da autora (2020).

Em seguida, como mostra a Figura 23, foi criada a base do mapa selecionando uma localização padrão, para que ao abrirmos o mapa, ele já esteja mostrando a cidade de Lajeado. Na sequência, foi implementado o mapa com as informações de longitude e latitude, criando ainda um arquivo HTML, onde o mapa de calor das ruas de Lajeado é apresentado usando serviços do OpenStreetMap.[1]

Figura 23 - Continuação do código para a obtenção da geolocalização

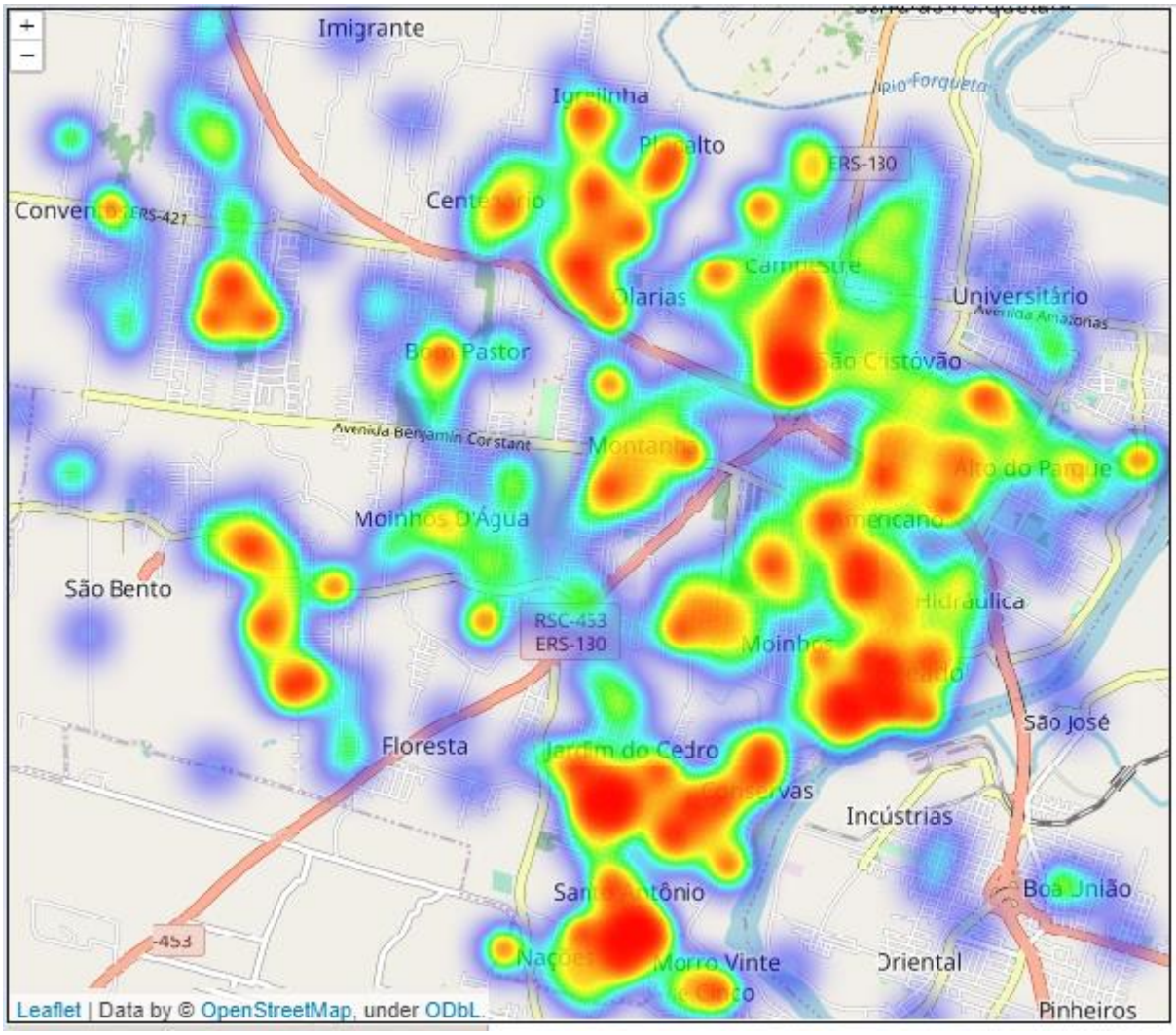
```
def generateBaseMap(default_location=[-29.50, -52.01], default_zoom_start=12):
    base_map = fl.Map(location=default_location, control_scale=True, zoom_start=default_zoom_start)
    return base_map

lat,lon=zip(*np.array(district_locations['geo_loc']))
district_locations['lat']=lat
district_locations['lon']=lon
basemap=generateBaseMap()
HeatMap(district_locations[['lat','lon','count']].values.tolist(),zoom=25,radius=15).add_to(basemap)
basemap.save("mapa_cliente.html")
```

Fonte: Da autora (2020).

A Figura 24, representa a visualização do mapa disponível através da página HTML. Nota-se, que a intensidade das cores varia, conforme a densidade de registros por rua. Desta forma, ruas com maior número de registros, aparecem com pontos mais avermelhados e ruas com casos mais isolados de registros, aparecem com pontos em tons azulados. O mapa é apresentado, focando diretamente na cidade de Lajeado ao ser aberto, mas conta com o mapa mundi completo.

[1]O OpenStreetMap é desenvolvido usando dados abertos e implementado por grupos voluntários de mapeadores, que colaboram e mantém atualizados dados referentes montanhas, rios, divisões políticas, estradas e demais estruturas do mundo todo (OPENSTREETMAP, 2020).

Figura 24 - Mapa de calor das ruas de Lajeado



Fonte: Da autora (2020).

O mapa é completamente interativo, permitindo que sejam observadas as ruas de Lajeado de forma isolada, quando aproxima-se para verificar de qual rua se trata. Também é possível se afastar para observar as ruas de uma forma mais ampla, incluindo mais ruas ao mesmo tempo. O ganho desta opção é que ao diminuir o *zoom,* é possível identificar bairros com alto índice de violência, através das ruas próximas com grandes números de registros de violência.

Também é possível se mover pelo mapa, podendo pressionar os botões do teclado de setas para cima e para baixo ou para os lados direito e esquerdo ou através de cliques com o *mouse*. O *zoom* pode ainda ser feito através do *scroll* do *mouse*.

5.3 Análise de grupos

Nesta tarefa diversas tentativas de agrupamento foram realizadas, como:

- Agrupar faixa etária da vítima versus grau de escolaridade da vítima, para que se pudesse encontrar alguma similaridade entre as vítimas, referente às suas faixas etárias e seus níveis de escolaridade;

- Agrupar por faixa etária versus turno, para que se pudesse identificar se algum grupo de faixa etária estaria sofrendo violência em um turno que deveria estar na escola, por exemplo;

- Agrupar pelo local da violência versus bairro, com o objetivo de agrupar os bairros mais violentos versus o local destas violências (residência da vítima, via pública, bares);

- Agrupar bairro versus dia da semana, com o intuito de agrupar os bairros violentos e os dias da semana com mais registros de violência, para indicar uma possível rota para policiamento em determinados bairros por dia da semana.

No entanto, estas análises não foram bem sucedidas, pois os dados utilizados são em sua maioria discretos ou cadastrados com códigos sequenciais, em que não há relação direta com valores anteriores ou posteriores.

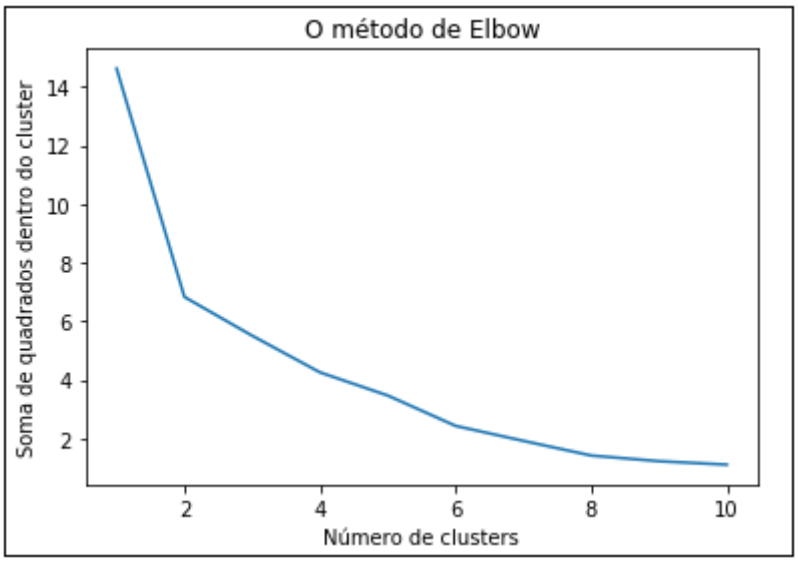
Por último, com as coordenadas de longitude e latitude, obtidas com o código do *notebook* do Mapa de Calor, foi possível gerar um agrupamento dos bairros através das suas coordenadas geográficas. Na primeira tentativa de execução, foram identificadas no gráfico gerado pelo algoritmo K-Means, coordenadas bem distantes que pareciam não pertencer a Lajeado. Então, estas coordenadas foram analisadas na base de dados.

Foi identificado que estas coordenadas pertenciam a ruas nomeadas de “INTERIOR” e “VIA PÚBLICA”, desta forma, as coordenadas retornadas eram inválidas e foram classificadas como *outliers*. Foi necessário voltar à etapa de

preparação dos dados para eliminar estes *outliers*, sendo gerada uma nova base final com os dados corretos. O mapa de calor também foi novamente processado para criar a base com as coordenadas corretas. Ao final, o algoritmo K-Means foi novamente executado.

No *notebook* criado, as bibliotecas relacionadas foram importadas e a base com as coordenadas também. Após, foi criado um vetor com as coordenadas e o método Elbow foi aplicado. Com um *range* de um até dez *clusters*, o método percorreu todo o vetor, calculando o benefício para cada quantidade de *clusters*. Como exibido na Figura 25, concluiu-se por meio do gráfico gerado pelo método Elbow, que neste caso o melhor número para a quantidade de *clusters* é seis.

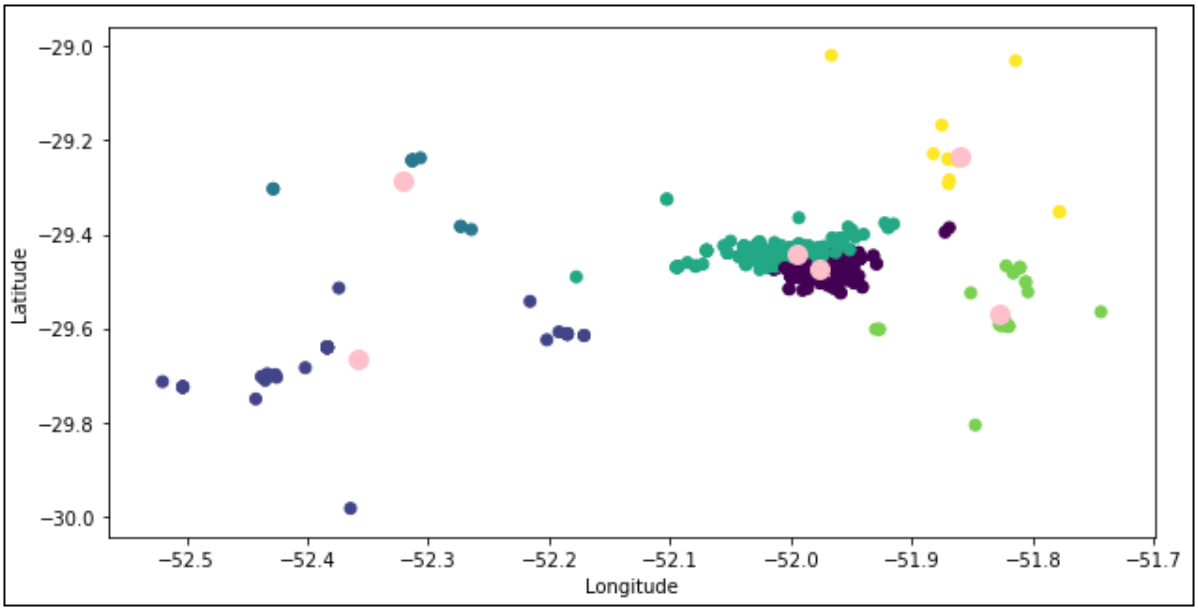
Figura 25- Gráfico do método Elbow



Fonte: Da autora (2020).

Com o número de *clusters* definido, foi possível construir o modelo com o algoritmo K-Means, computar os centróides para cada *cluster* e prever o índice de *cluster* para cada amostra. Então foi possível plotar os *clusters* e seus respectivos centróides com diferentes cores, para facilitar a visualização, conforme a Figura 26.

Figura 26 - *Clusters* de latitude e longitude



Fonte: Da autora (2020).

Os números de *clusters* apresentado na Figura 26, estão diretamente relacionados com os pontos de calor, apresentados no mapa de calor, na Figura 24. No mapa de calor, já era possível identificar alguns pontos da cidade com maior intensidade de violência e o agrupamento do K-Means torna isso também muito visível.

5.4 Associação

Nesta tarefa, primeiramente foram utilizados somente os atributos de violências e agressões, para que se pudessem encontrar associações entre os atributos, conforme os registros. Foi identificado que quando há violência psicológica, há fortes chances de a violência física ser consequente. Posteriormente, foram adicionados os atributos de ciclo de vida e de sexo, tanto das vítimas como dos autores das violências. Foi então identificado um melhor resultado, que será apresentado a seguir.

Para a utilização do algoritmo Apriori, tem-se um melhor comportamento quando os atributos estão em forma de valores binários. Deste modo, foi necessário retornar para a etapa de preparação dos dados e separar os objetos do ciclo de vida e do sexo do autor para torná-los atributos binários. Como a vítima não possuía ciclo

de vida, mas apenas faixa etária, foram criados atributos novos, referente aos ciclos de vida da mesma, baseados na idade das vítimas. Desta forma, também foram criados atributos binários do ciclo de vida e sexo das vítimas.

Com os dados preparados para serem processados no algoritmo, um novo notebook foi criado, no qual as bibliotecas relacionadas e a base de dados foram importadas. A seguir, um Data Frame foi criado apenas com os atributos binários e necessários para aplicar a técnica de associação. Em seguida, foi criado um código para encontrar todos os grupos de itens frequentes. Para isso, foi definido o valor mínimo para o suporte de 30%, como é apresentado na Figura 27.

Figura 27 - Código para encontrar os grupos de itens frequentes

```
frequent_itemsets = apriori(df[['VIOL_FISIC,C,1', 'VIOL_PSICO,C,1
         'vitima_adolescente','vitima_jovem','vitima_adulto','vit
         'autora_feminina','autor_ambos']],
                             min_support=0.30, use_colnames=True)
```

Fonte: Da autora (2020).

Foram encontrados 37 conjuntos de itens frequentes com suporte superior a 30%, sendo que sete deles são apresentados na Figura 28.

Figura 28 - Grupos de itens frequentes

	support	itemsets
0	0.756537	(VIOL_FISIC,C,1)
1	0.694814	(VIOL_PSICO,C,1)
2	0.525075	(vitima_adulto)
3	0.503643	(autor_adulto)
4	0.773253	(vitima_feminina)
5	0.740677	(autor_masculino)
6	0.537077	(VIOL_PSICO,C,1, VIOL_FISIC,C,1)
7	0.417488	(vitima_adulto, VIOL_FISIC,C,1)

Fonte: Da autora (2020).

Na etapa final, foram criadas as regras de associação, a partir dos grupos de itens frequentes, utilizando a métrica *lift*. Dada a quantidade de regras formadas, foi estabelecido um filtro para que apenas regras com *lift* maior que 1.125 e confiança

maior que 0.85 fossem consideradas. A seguir, é apresentada a Figura 29, com as regras consideradas válidas, visto que atingem as medidas de interesse

Figura 29 - Regras de associação de objetos

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(vitima_adulto, VIOL_FISIC,C,1)	(vitima_feminina)	0.417488	0.773253	0.363480	0.870637	1.125940	0.040656	1.752788
(vitima_adulto, VIOL_PSICO,C,1)	(vitima_feminina)	0.375054	0.773253	0.339906	0.906286	1.172042	0.049894	2.419553
(vitima_adulto, VIOL_PSICO,C,1)	(autor_masculino)	0.375054	0.740677	0.320617	0.854857	1.154156	0.042824	1.786673
(vitima_adulto, vitima_feminina)	(autor_masculino)	0.459923	0.740677	0.391342	0.850885	1.148794	0.050687	1.739083
(autor_adulto, autor_masculino)	(vitima_feminina)	0.417060	0.773253	0.370767	0.889003	1.149692	0.048275	2.042820
(VIOL_FISIC,C,1, vitima_feminina, VIOL_PSICO,C,1)	(autor_masculino)	0.419631	0.740677	0.358766	0.854954	1.154287	0.047954	1.787866
(vitima_adulto, vitima_feminina, VIOL_FISIC,C,1)	(autor_masculino)	0.363480	0.740677	0.310330	0.853774	1.152693	0.041108	1.773433
(vitima_adulto, vitima_feminina, VIOL_PSICO,C,1)	(autor_masculino)	0.339906	0.740677	0.300043	0.882724	1.191779	0.048282	2.211215

Fonte: Da autora (2020).

A Figura 29, apresenta uma associação de regras de interferência estatística, exclusivamente com base nos dados utilizados. Para a validação dessas regras, caberia somente aos especialistas da área da saúde fazerem uma análise qualitativa. No entanto, será apresentado uma explicação apenas para interpretação da Figura 29.

Analisando a primeira regra é indicado, que quando há um registro contendo antecedentes com vítimas adultas e violência físicas, há como consequente, uma vítima feminina. Sendo que, o seu suporte possui uma frequência de 36% do evento na base de dados e uma confiança de 87%, entre os antecedentes e consequente. O *lift* apresenta uma probabilidade de 1.26 da vítima ser feminina, quando os antecedentes forem vítimas adultas e violências físicas. Já a *leverage* de 0.04 implica na dependência entre o antecedente e consequente. Por último, a convicção indica uma relação de 1.75 entre eles.

As regras com melhores valores e que são bem semelhantes entre si, são a quarta e quinta regras. Onde em uma, há como antecedentes, uma vítima adulta e feminina e como consequente, o autor da violência masculino e na outra, há um antecedente, como autor adulto e masculino, e como consequência, uma vítima feminina. Ambos eventos são frequentes na base de dados e possuem uma confiança alta.

A sexta regra tem uma particularidade, devido aos antecedentes serem as violências física e psicológicas contra vítimas femininas e o consequente ser um autor masculino. Também há uma frequência relevante deste evento na base e uma alta convicção.

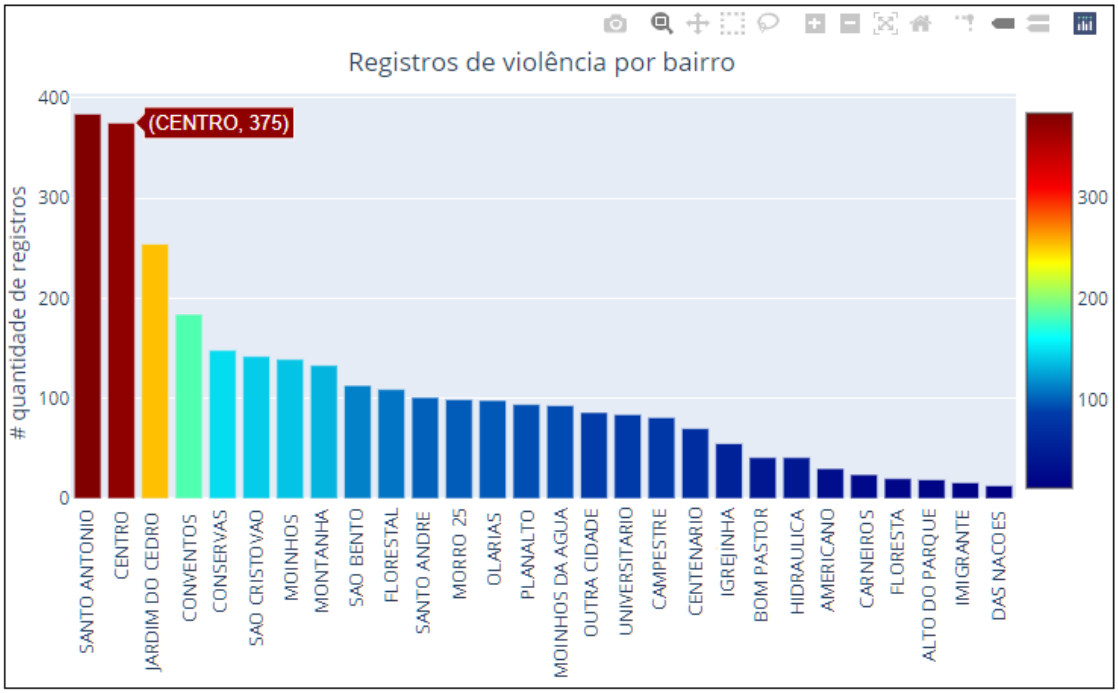
5.5 Apresentação para os gestores

Todos os gráficos gerados na etapa de modelagem foram reproduzidos de forma mais simples e interativa, para facilitar o entendimento por parte da área de gestão. Alguns exemplos serão apresentados a seguir.

O primeiro gráfico representa o índice de notificação de violência, agrupado por bairros, conforme foi apresentado na Figura 13. Nesta etapa, o gráfico gerado conta com um ícone de ‘câmera’ que possibilita fazer *download* do gráfico. Na lupa é possível dar *zoom,* enquadrando especificamente a parte que se deseja aproximar e analisar mais detalhadamente. Também estão presentes de outras opções, como uso de auto escala para voltar à escala original.

Passando o *mouse* por cima da barra de cada bairro, é possível ler o nome do bairro em questão e identificar o número exato de casos registrados para aquele bairro. Ao lado direito, também foi inserida uma barra de cores, que define a cor dos bairros, de acordo com o seu número de registros. Na Figura 30 é possível visualizar que, por exemplo, o bairro CENTRO possui exatamente 375 registros de violência, estando na cor vermelho forte. Conforme a barra de escala de cores, esta cor representa bairros com mais de 300 registros.

Figura 30 - Gráfico interativo do agrupamento dos bairros

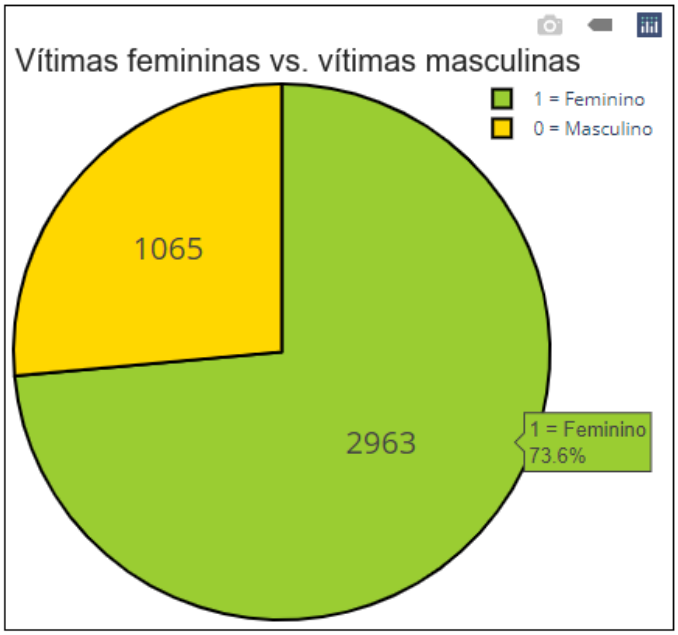


Fonte: Da autora (2020).

Na Figura 31, é apresentado o gráfico de vítimas femininas versus vítimas masculinas, que foi apresentado na Figura 16. No modelo interativo, também é possível fazer o *download,* usando o ícone de ‘câmera’. Ao passar o *mouse* por cima de cada objeto, é possível visualizar a porcentagem de registros que cada sexo apresenta na base, sendo que o número exato de registros é apresentado de maneira fixa no gráfico, independente de interação humana.

Desta forma, podemos visualizar na Figura 31, que o sexo feminino representa exatamente 73,6% das vítimas, além de que foram registrados 2963 casos de violências contra mulheres e 1065 casos de violências contra homens.

Figura 31 - Gráfico interativo dos agrupamentos de vítimas femininas vs. masculinas

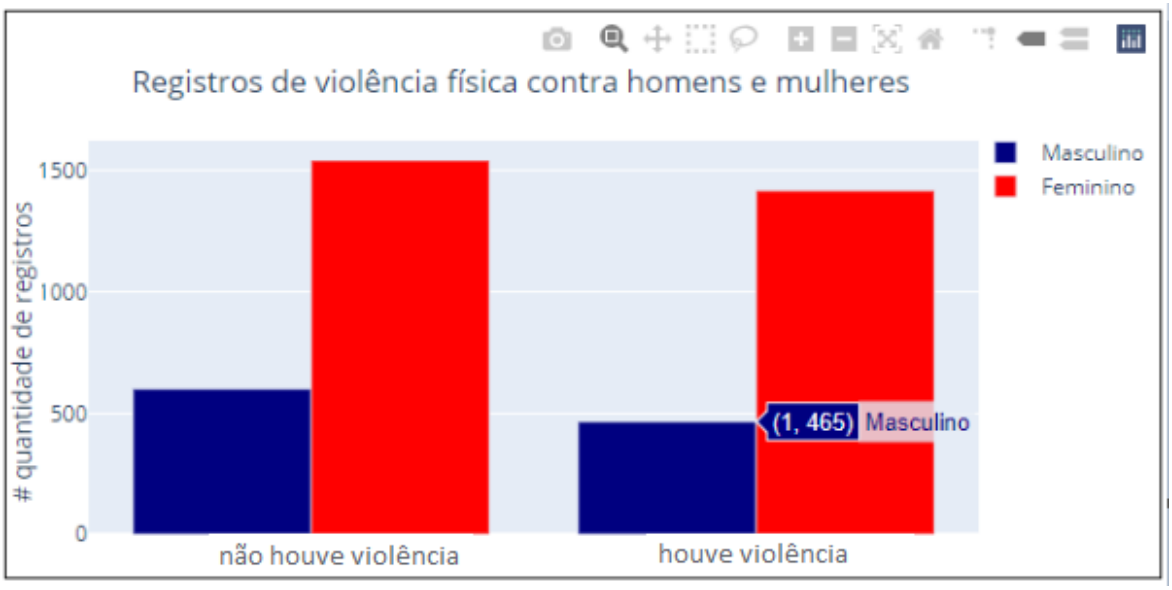


Fonte: Da autora (2020).

Na Figura 32, é apresentado o gráfico dos registros de violência física contra homens e mulheres, onde na etapa de modelagem na Figura 19, foram apresentados dois gráficos separados: um para registros de violência contra homens e o outro para registros de violência contra mulheres. Nesta etapa, a Figura 32, apresenta em vermelho, os registros de violência física contra mulheres e em azul, os registros de violência física contra homens. Estando lado a lado, para que além da facilidade da visualização, também seja possível comparar os números de registros de violência contra homens e contra mulheres.

No gráfico apresentado na Figura 32, é possível facilmente visualizar os casos onde houve violência física e pode-se verificar que na base de dados utilizada, há 1465 registros de violência física contra homens.

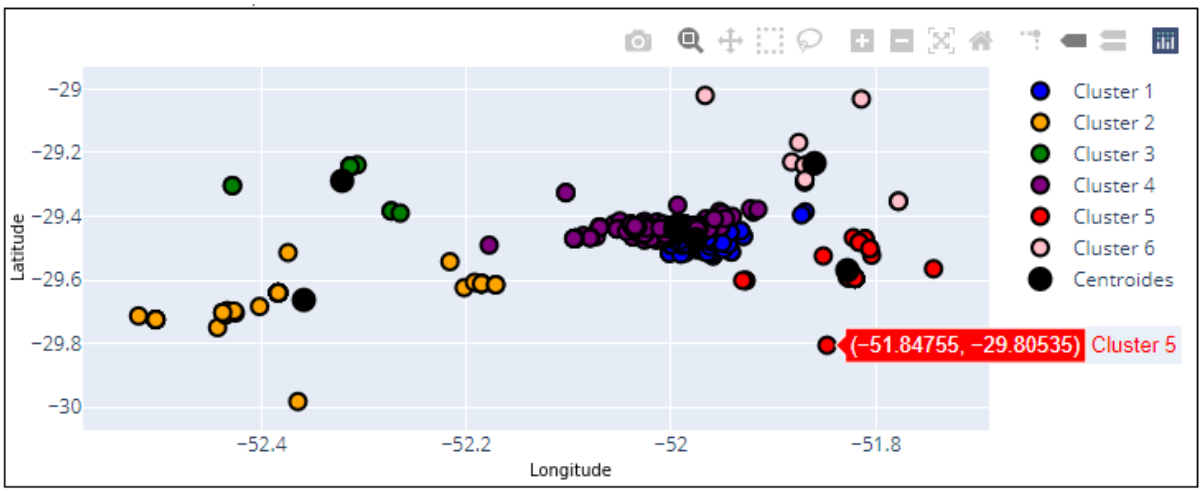
Figura 32 - Gráfico interativo dos registros de violências físicas contra homens e mulheres



Fonte: Da autora (2020).

Na Figura 33, é apresentado o gráfico interativo da tarefa de agrupamento, onde são apresentados os *clusters* com cores diferentes, também permitindo fazer download do gráfico, além de permitir zoom, com enquadramento especificamente na parte que se deseja aproximar e analisar mais detalhadamente. Também há outras opções, como o botão para voltar à escala original. Passando o mouse por cima de cada ponto, é possível visualizar as coordenadas que ele representa e verificar a qual *cluster* ele pertence.

Figura 33 - Gráfico interativo do agrupamento de coordenadas



Fonte: Da autora (2020).

O mapa de calor implementado em formato HTML, já é desde seu princípio interativo e de fácil manuseio, desta forma não precisando sofrer alteração nesta etapa.

6 CONCLUSÃO

Devido à aplicação de algoritmos de Mineração de Dados, hoje é possível extrair conhecimento de grande valor para as organizações. Com base no trabalho realizado e nos resultados alcançados, ficou demonstrado que, através de técnicas especializadas, é possível explorar grandes volumes de dados e encontrar padrões de interesse.

Também é viável descobrir regras e relações ocultas, com o potencial de auxiliar gestores no processo de tomada de decisões. No entanto, foi necessária uma sistematização do fluxo do processo de Análise e Mineração de Dados, pois o entendimento do negócio e dos dados, bem como a escolha correta das ferramentas e validação dos resultados, são etapas cruciais, para que novas informações úteis e até então desconhecidas sejam encontradas.

Para compreensão do caso estudado, contatos foram feitos com o projeto Pacto Lajeado pela Paz, onde foi possível perceber o entusiasmo das partes envolvidas, devido as possibilidades da extração de conhecimento. A autora reitera seu agradecimento, pelo apoio recebido de todos os envolvidos. Contudo, devido à pandemia do ano de 2020, o presente trabalho não pôde contar com a análise especializada. Todavia, foi possível realizar um trabalho que alcançou os objetivos, apresentando análise descritiva, análise de grupos e associação de regras.

Foi possível visualizar informações consideradas relevantes, como a faixa etária dos possíveis praticantes de ações violentas, o número de notificações por bairro, o acentuado número de violência contra mulheres e a observação de que

grande parte da violência é causada pelo sexismo. Também foram descobertas algumas regras de associação que trazem, por exemplo, que onde há antecedentes de violência física e psicológica contra uma vítima feminina, o consequente é um novo caso de violência por autor de sexo masculino.

Objetivou-se implantar um fluxo do processo de Análise e Mineração de Dados na Prefeitura de Lajeado, auxiliando na estruturação da base de dados e na criação de modelos que possam identificar padrões, relações, anomalias e regras ocultas nos dados brutos originários da Secretaria da Saúde. Objetivou-se também entregar estes modelos em um formato que usuários fora da área de Tecnologia da Informação pudessem repetir o processo com novos dados futuramente, daí o estabelecimento de um fluxo de dados e análises via *notebooks.*

Após o desenvolvimento do presente trabalho, algumas propostas de melhorias e também sugestões para projetos futuros foram identificadas. Dentre elas, o acesso à análise especializada pela Secretaria da Saúde, que estava sendo proposto no início do trabalho, parte integrante do fluxo de Análise e Mineração de Dados. Também seria interessante trabalhar com os dados da relação entre a vítima e o agressor.

Outro ponto, seria realizar a exploração dos dados das demais secretarias envolvidas no projeto, Pacto Lajeado pela Paz, possibilitando o cruzamento entre bases de dados. Além disso, poderiam ser treinados modelos preditivos, utilizando exemplos de casos de reincidência de violência, para inferir casos de violência que tendem a ser reincidentes e tentar preveni-los.

REFERÊNCIAS

ANACONDA, Anaconda Distribution. Texas, 2019. Disponível em:<https://www.anaconda.com/distribution/>. Acesso em: 02 nov. 2019.

AZEVEDO, Ana Isabel Rojão Lourenço; SANTOS, Manuel Filipe. **KDD, SEMMA and CRISP-DM**: a parallel overview. IADS-DM, 2008. Disponível em: <https://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf>. Acesso em: 02 ago. 2019.

BERRY, Michael J. A.; LINOFF, Gordon. **Data mining techniques**. USA : Wiley Computer Publishing, 1997.

BORGES, Luiz Eduardo Borges. **Python para Desenvolvedores**. 2. ed. Rio de Janeiro, 2010. E-book. Disponível em: <https://ricardoduarte.github.io/python-para-desenvolvedores/download/python_para_desenvolvedores_2ed.pdf>. Acesso em: 09 out. 2019.

BRAZ, Lucas; FERREIRA, Rafael; DERMEVAL, Diego; VERAS, Douglas. **Aplicando Mineração de Dados para Apoiar na Tomada de Decisão na Segurança Pública no Estado de Alagoas**. Universidade Federal de Alagoas, 2009. Disponível em: <https://www.researchgate.net/publication/233843302_Aplicando_Mineracao_de_Dados_para_Apoiar_na_Tomada_de_Decisao_na_Seguranca_Publica_no_Estado_de_Alagoas>. Acesso em 29 ago. 2019.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: **Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), p. 1-29, 2009. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf> . Acesso em: 04 set. 2019.

CARVALHO, Luís Alfredo Vidal. **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração** - Ciência Moderna - RJ, 2005.

CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à mineração de dados**. Editora Saraiva, São Paulo, 2016.

CERVO, Amado L.; BERVIAN, Pedro A.; DA SILVA, Roberto. **Metodologia Científica**. 6. Ed. São Paulo: Pearson Prentice Hall: 2007. E-book. Disponível em: <http://www.univates.br/biblioteca>. Acesso em 08 out. 2019.

CHAPMAN, Pete et al. CRISP-DM 1.0: **Step-by-step data mining guide**. SPSS inc, 2000. Disponível em: <https://www.the-modeling-agency.com/crisp-dm.pdf>. Acesso em: 02 set. 2019.

COPPIN, Ben**. Inteligência Artificial**. Rio de Janeiro: LTC, 2013.

DA SILVA, Fabricio Machado; LENZ, Maikon Lucian; FREITAS Pedro Henrique; DOS SANTOS, Sydnei Cerqueira. **Inteligência Artificial**. Porto Alegre: SAGAH, 2019.

DE AMO, Sandra; **Técnicas de mineração de dados**. Universidade Federal de Uberlândia, 2004.

DIAS, Maria Madalena. **Parâmetros na escolha de técnicas e ferramentas de mineração de dados**. Acta Scientiarum. Technology, v. 24, p. 1715-1725, 2002.

FACELI Katti; LORENA Ana Carolina; GAMA João; CARVALHO André. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Editorial Nacional, 2011.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases**. AI magazine, v. 17, n. 3, p. 37, 1996.

FERNANDES, Anita Maria da Rocha. **Inteligência artificial: Noções gerais**, Visual Books, 2008.

GANDHI, Pryha; SHARMA, Shayog; **Review of Predictive Modeling on Crime Pattern against Women**. International Journal of Recent Research Aspects, 2017. Disponivel em: <http://web.a.ebscohost.com/ehost/detail/detail?vid=0&sid=49508639-36d2-4137-9e0f-304a782f24d8%40sessionmgr4007&bdata=Jmxhbmc9cHQtYnlmc2l0ZT1laG9zdC1saXZl#AN=129311385&db=afh>. Acesso em: 28 ago. 2019.

GEOPY, 2018. **Documentation manual**. Dispoível em: <https://geopy.readthedocs.io/en/stable/#module-geopy.geocoders> Acesso em 03 junho 2020.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques**. University of Ilinois at Urbana-Champaign: Elsevier, 2006.

HAN Jiawei; KAMBER Micheline; PEI Jian. **Data Mining: Concepts and Techniques, 3º ed**. Morgan Kaufmann Publishers, 2012.

JAYAWEERA, Isuru; SAJEEWA, Chamath; LIYANAGE, Sampath; WIJEWARDANE, Tharindu; PERERA, Indika; WIJAYASIRI, Adeesha. **Crime analytics: Analysis of crimes through newspaper articles**. Moratuwa Engineering Research Conference (MERCon), 2015. Moratuwa, Sri Lanka, p. 277–282, 2015. Disponível em: <https://www.researchgate.net/publication/275950527_Crime_Analytics_Analysis_of_Crimes_Through_Newspaper_Articles>. Acesso em: 28 ago. 2019.

KAUARK, Fabiana da Silva; MANHÃES, Fernanda Castro; MEDEIROS, Carlos Henrique. **Metodologia da Pesquisa**: Um guia prático. Itabuna, 2010.

KEYVANPOURA, Mohammad R.; JAVIDEHB, Mostafa; EBRAHIMI, Mohammad R. **Detecting and investigating crime by means of data mining: a general crime matching framework**. Procedia Computer Science, Elsevier, v. 3, p. 872–880, 2011. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050910005181>. Acesso em 20 de ago. 2019.

KIMBALL, Ralph; CASERTA, Joe. **The data warehouse ETL toolkit**: practical techniques for extracting, cleaning, conforming and delivering data. Indianapolis: Wiley, 2004.

LAJEADO, Prefeitura Municipal. Disponivel em: <https://www.lajeado.rs.gov.br/?titulo=Pacto%20Lajeado%20pela%20Paz&template=conteudo&categoria=1027&codigoCategoria=1027> Acesso em 10 out. 2019..

LAROSE, Daniel; LAROSE, Chantal. **Discovering knowledge in data: an introduction to data mining**. John Wiley & Sons, 2014.

MCKINNEY, Wes. **Python for data analysis**: Data wrangling with Pandas, NumPy, and IPython. 1. ed. Sebastopol: O'Reilly Media, 2012.

NETO, Silvino D. **Mineração de dados de ocorrências criminais para identificação de zonas de alta criminalidade em Fortaleza e região metropolitana**. Universidade Federal do Ceará: Campus de Quixadá, 2017. Disponível em: <www.repositorio.ufc.br/bitstream/riufc/29565/1/2017_tcc_sdneto.pdf>. Acesso em 18 de ago. 2019.

OPENSTREETMAP, 2020. Disponivel em: <https://www.openstreetmap.org/about>. Acesso em 07 jun. 2020

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent; VANDERPLAS, Jake; PASSOS, Alexandre; COURNAPEAU, David; BRUCHER, Matthieu; PERROT, Matthieu; DUCHESNAY, Édouard. **Scikit-learn: machine learning in Python**. Journal of Machine Learning Research, n.12, p. 2825-2830, 2011. Disponível em

<http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf> Acesso em 02 jun. 2020.

PLOTLY, 2020. Disponivel em: <https://plotly.com/python/getting-started/>Acesso em 03 jun. 2020.

PRODANOV, Cleber C.; DE FREITAS, Ernani C.; Metodologia do Trabalho Científico: **Métodos e Técnicas da Pesquisa e do trabalho Acadêmico**. 2. ed. Novo Hamburgo: Feevale, 2013. E-book. Disponível em: <http://www.feevale.br/Comum/midias/8807f05a-14d0-4d5b-b1ad-1538f3aef538/E-book%20Metodologia%20do%20Trabalho%20Cientifico.pdf> Acessado em: 08 out. 2019.

REZENDE, Solange Oliveira; PUGLIESI Jaqueline; MELANDA Edson; DE PAULA Marcos. **Mineração de Dados**, in REZENDE, Solange Oliveira (Eds), Sistemas inteligentes. Editora Manole Ltda., p.307-335. 2003.

REZENDE, Solange Oliveira. **Mineração de Dados**. UNISINOS, São Leopoldo, RS, 2004.

SHARMA, Anubha; TIWARI, Nirupama; Efficient Fuzzy **Apriori Association Rule Mining to find Co-occurrence Relationship**. India, 2014. Disponível em: <http://www.iasir.net/IJSWSpapers/IJSWS14-147.pdf> Acesso em:10 jun. 2020.

SILVA, Edilberto Magalhães; **Descoberta de conhecimento com o uso de Text Mining:** Cruzando o abismo de Moore. Universidade Católica de Brasilia, 2002. Disponível em: <https://bdtd.ucb.br:8443/jspui/bitstream/123456789/1462/1/Dissertacao_Edilberto.pdf> Acesso em: 28 set. 2019.

UBER, José Lino. **Descoberta de Conhecimento com o uso de Text Mining aplicada ao SAC**. Universidade Regional de Blumenau, 2004. Disponível em: <http://dsc.inf.furb.br/arquivos/tccs/monografias/2004-2joselubervf.pdf>. Acesso em: 11 out. 2019.

VARELLA, Jorge Luis; QUADRELLI, Giovane**; Redes Neurais e análise de potência. Revista de Tecnologia Aplicada** (RTA), v.6, n.3, p.33-45. 2017.

WHITBY, Blay. **Inteligência artificial**: um guia para iniciantes. One World Publications, 2003.

WIRTH, Rüdiger; HIPP, Jochen. **CRISP-DM**: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Citeseer, 2000. Disponível em: <https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining>. Acessado em: 07 out. 2019.

ANEXOS

ANEXO A - Formulário SINAN

República Federativa do Brasil
Ministério da Saúde

SINAN
SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO
FICHA DE NOTIFICAÇÃO INDIVIDUAL

Nº

Caso suspeito ou confirmado de violência doméstica/intrafamiliar, sexual, autoprovocada, tráfico de pessoas, trabalho escravo, trabalho infantil, tortura, intervenção legal e violências homofóbicas contra mulheres e homens em todas as idades. No caso de violência extrafamiliar/comunitária, somente serão objetos de notificação as violências contra crianças, adolescentes, mulheres, pessoas idosas, pessoa com deficiência, indígenas e população LGBT.

Dados Gerais / Notificação Individual

1 Tipo de Notificação — 2 - Individual

2 Agravo/doença — VIOLÊNCIA INTERPESSOAL/AUTOPROVOCADA — Código (CID10) Y09 — 3 Data da notificação

4 UF — 5 Município de notificação — Código (IBGE)

6 Unidade Notificadora — 1- Unidade de Saúde 2- Unidade de Assistência Social 3- Estabelecimento de Ensino 4- Conselho Tutelar 5- Unidade de Saúde Indígena 6- Centro Especializado de Atendimento à Mulher 7- Outros

7 Nome da Unidade Notificadora — Código Unidade — 9 Data da ocorrência da violência

8 Unidade de Saúde — Código (CNES)

10 Nome do paciente — 11 Data de nascimento

12 (ou) Idade — 1 - Hora 2 - Dia 3 - Mês 4 - Ano — 13 Sexo M - Masculino F - Feminino I - Ignorado — 14 Gestante 1-1ºTrimestre 2-2ºTrimestre 3-3ºTrimestre 4- Idade gestacional ignorada 5-Não 6- Não se aplica 9-Ignorado — 15 Raça/Cor 1-Branca 2-Preta 3-Amarela 4-Parda 5-Indígena 9- Ignorado

16 Escolaridade
0-Analfabeto 1-1ª a 4ª série incompleta do EF (antigo primário ou 1º grau) 2-4ª série completa do EF (antigo primário ou 1º grau) 3-5ª à 8ª série incompleta do EF (antigo ginásio ou 1º grau) 4-Ensino fundamental completo (antigo ginásio ou 1º grau) 5-Ensino médio incompleto (antigo colegial ou 2º grau) 6-Ensino médio completo (antigo colegial ou 2º grau) 7-Educação superior incompleta 8-Educação superior completa 9-Ignorado 10- Não se aplica

17 Número do Cartão SUS — 18 Nome da mãe

Dados de Residência

19 UF — 20 Município de Residência — Código (IBGE) — 21 Distrito

22 Bairro — 23 Logradouro (rua, avenida,...) — Código

24 Número — 25 Complemento (apto., casa, ...) — 26 Geo campo 1

27 Geo campo 2 — 28 Ponto de Referência — 29 CEP

30 (DDD) Telefone — 31 Zona 1 - Urbana 2 - Rural 3 - Periurbana 9 - Ignorado — 32 País (se residente fora do Brasil)

Dados Complementares

Dados da Pessoa Atendida

33 Nome Social — 34 Ocupação

35 Situação conjugal / Estado civil
1 - Solteiro 2 - Casado/união consensual 3 - Viúvo 4 - Separado 8 - Não se aplica 9 - Ignorado

36 Orientação Sexual
1-Heterossexual 2-Homossexual (gay/lésbica) 3-Bissexual 8-Não se aplica 9-Ignorado

37 Identidade de gênero:
1-Travesti 2-Mulher Transexual 3-Homem Transexual 8-Não se aplica 9-Ignorado

38 Possui algum tipo de deficiência/ transtorno?
1- Sim 2- Não 9- Ignorado

39 Se sim, qual tipo de deficiência /transtorno? 1- Sim 2- Não 8-Não se aplica 9- Ignorado
☐ Deficiência Física ☐ Deficiência visual ☐ Transtorno mental ☐ Outras
☐ Deficiência Intelectual ☐ Deficiência auditiva ☐ Transtorno de comportamento

Dados da Ocorrência

40 UF — 41 Município de ocorrência — Código (IBGE) — 42 Distrito

43 Bairro — 44 Logradouro (rua, avenida,...) — Código

45 Número — 46 Complemento (apto., casa, ...) — 47 Geo campo 3 — 48 Geo campo 4

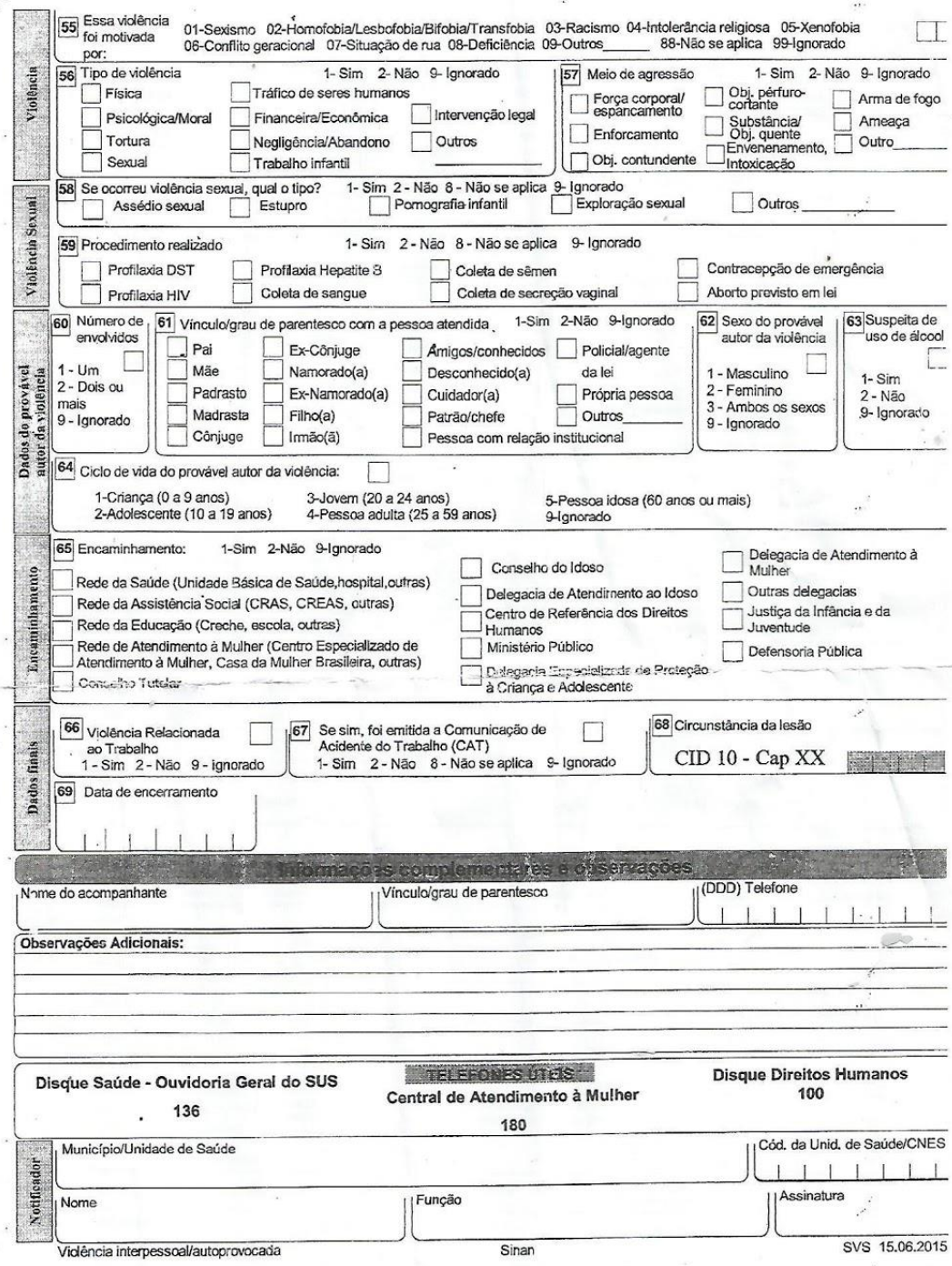
49 Ponto de Referência — 50 Zona 1 - Urbana 2 - Rural 3 - Periurbana 9 - Ignorado — 51 Hora da ocorrência (00:00 - 23:59 horas)

52 Local de ocorrência
01 - Residência 02 - Habitação coletiva 03 - Escola 04 - Local de prática esportiva 05 - Bar ou similar 06 - Via pública 07 - Comércio/serviços 08 - Indústrias/construção 09 - Outro _____ 99 - Ignorado

53 Ocorreu outras vezes? 1 - Sim 2 - Não 9 - Ignorado

54 A lesão foi autoprovocada? 1 - Sim 2 - Não 9 - Ignorado

SVS 15.06.2015



Violência

55 Essa violência foi motivada por: 01-Sexismo 02-Homofobia/Lesbofobia/Bifobia/Transfobia 03-Racismo 04-Intolerância religiosa 05-Xenofobia 06-Conflito geracional 07-Situação de rua 08-Deficiência 09-Outros_____ 88-Não se aplica 99-Ignorado

56 Tipo de violência — 1- Sim 2- Não 9- Ignorado
- ☐ Física
- ☐ Psicológica/Moral
- ☐ Tortura
- ☐ Sexual
- ☐ Tráfico de seres humanos
- ☐ Financeira/Econômica
- ☐ Negligência/Abandono
- ☐ Trabalho infantil
- ☐ Intervenção legal
- ☐ Outros _____

57 Meio de agressão — 1- Sim 2- Não 9- Ignorado
- ☐ Força corporal/ espancamento
- ☐ Enforcamento
- ☐ Obj. contundente
- ☐ Obj. pérfuro-cortante
- ☐ Substância/ Obj. quente
- ☐ Envenenamento, Intoxicação
- ☐ Arma de fogo
- ☐ Ameaça
- ☐ Outro_____

Violência Sexual

58 Se ocorreu violência sexual, qual o tipo? 1- Sim 2 - Não 8 - Não se aplica 9- Ignorado
☐ Assédio sexual ☐ Estupro ☐ Pornografia infantil ☐ Exploração sexual ☐ Outros _____

59 Procedimento realizado — 1- Sim 2 - Não 8 - Não se aplica 9- Ignorado
- ☐ Profilaxia DST
- ☐ Profilaxia HIV
- ☐ Profilaxia Hepatite B
- ☐ Coleta de sangue
- ☐ Coleta de sêmen
- ☐ Coleta de secreção vaginal
- ☐ Contracepção de emergência
- ☐ Aborto previsto em lei

Dados do provável autor da violência

60 Número de envolvidos ☐
1 - Um
2 - Dois ou mais
9 - Ignorado

61 Vínculo/grau de parentesco com a pessoa atendida — 1-Sim 2-Não 9-Ignorado
- ☐ Pai
- ☐ Mãe
- ☐ Padrasto
- ☐ Madrasta
- ☐ Cônjuge
- ☐ Ex-Cônjuge
- ☐ Namorado(a)
- ☐ Ex-Namorado(a)
- ☐ Filho(a)
- ☐ Irmão(ã)
- ☐ Amigos/conhecidos
- ☐ Desconhecido(a)
- ☐ Cuidador(a)
- ☐ Patrão/chefe
- ☐ Pessoa com relação institucional
- ☐ Policial/agente da lei
- ☐ Própria pessoa
- ☐ Outros_____

62 Sexo do provável autor da violência ☐
1 - Masculino
2 - Feminino
3 - Ambos os sexos
9 - Ignorado

63 Suspeita de uso de álcool ☐
1- Sim
2 - Não
9- Ignorado

64 Ciclo de vida do provável autor da violência: ☐
1-Criança (0 a 9 anos)
2-Adolescente (10 a 19 anos)
3-Jovem (20 a 24 anos)
4-Pessoa adulta (25 a 59 anos)
5-Pessoa idosa (60 anos ou mais)
9-Ignorado

Encaminhamento

65 Encaminhamento: 1-Sim 2-Não 9-Ignorado
- ☐ Rede da Saúde (Unidade Básica de Saúde,hospital,outras)
- ☐ Rede da Assistência Social (CRAS, CREAS, outras)
- ☐ Rede da Educação (Creche, escola, outras)
- ☐ Rede de Atendimento à Mulher (Centro Especializado de Atendimento à Mulher, Casa da Mulher Brasileira, outras)
- ☐ Conselho Tutelar
- ☐ Conselho do Idoso
- ☐ Delegacia de Atendimento ao Idoso
- ☐ Centro de Referência dos Direitos Humanos
- ☐ Ministério Público
- ☐ Delegacia Especializada de Proteção à Criança e Adolescente
- ☐ Delegacia de Atendimento à Mulher
- ☐ Outras delegacias
- ☐ Justiça da Infância e da Juventude
- ☐ Defensoria Pública

Dados finais

66 Violência Relacionada ao Trabalho ☐
1 - Sim 2 - Não 9 - ignorado

67 Se sim, foi emitida a Comunicação de Acidente do Trabalho (CAT) ☐
1- Sim 2 - Não 8 - Não se aplica 9- Ignorado

68 Circunstância da lesão
CID 10 - Cap XX

69 Data de encerramento

Informações complementares e observações

Nome do acompanhante | Vínculo/grau de parentesco | (DDD) Telefone

Observações Adicionais:

TELEFONES ÚTEIS

Disque Saúde - Ouvidoria Geral do SUS
136

Central de Atendimento à Mulher
180

Disque Direitos Humanos
100

Notificador

Município/Unidade de Saúde | Cód. da Unid. de Saúde/CNES

Nome | Função | Assinatura

UNIVATES

R. Avelino Talini, 171 | Bairro Universitário | Lajeado | RS | Brasil
CEP 95914.014 | Cx. Postal 155 | Fone: (51) 3714.7000
www.univates.br | 0800 7 07 08 09