ANÁLISE DA VIOLÊNCIA CONTRA A MULHER COM MINERAÇÃO DE DADOS: UMA ABORDAGEM UTILIZANDO CLUSTERIZAÇÃO

ANALYSIS OF VIOLENCE AGAINST WOMEN WITH DATA MINING: AN APPROACH USING CLUSTERING

Andressa de Souza Santiago
https://orcid.org/0000-0003-3821-0081
450265928-21
Centro Paula Souza – Fatec Indaiatuba/SP
andressa.santiago01@fatec.sp.gov.br

Natalia Fiori dos Santos
https://orcid.org/0000-0002-4746-0966
465266608-00
Centro Paula Souza – Fatec Indaiatuba/SP
natalia.santos45@fatec.sp.gov.br

Orientador: Profa. Dra. Maria das Graças J. M. Tomazela
https://orcid.org/0000-0002-5471-2658
085107058-28
Centro Paula Souza – Fatec Sorocaba/SP
graca.tomazela@fatec.sp.gov.br

RESUMO: O Brasil é o 18º pais na América Latina a possuir uma lei para caso de violência doméstica contra a mulher, trata-se da Lei Maria da Penha. Os casos de violência contra mulher aumentaram mais de 150% no Brasil entre 2016 e 2017. Desta forma, o uso das tecnologias de informação pode contribuir para melhorar a eficiência da gestão desses dados. As técnicas de mineração de dados possibilitam a manipulação de grandes conjuntos de dados e a identificação de padrões novos nesses conjuntos. Assim o objetivo deste trabalho foi utilizar a abordagem experimental para analisar e caracterizar possíveis agrupamentos de forma a elucidar o comportamento dos casos de violência contra a mulher, contribuindo para a tomada de decisão nas ações e/ou políticas de combate à violência contra mulher. Inicialmente foram coletados dados do SINAN/DATASUS sobre Violência Doméstica, Sexual e outras violências sofridas pelo sexo feminino. Na sequência foram realizadas atividades de pré-processamento dos dados; aplicou-se então, por meio da ferramenta Weka, o método de clusterização para facilitar a compreensão dos fatores que influenciam na violência contra mulher. A análise dos clusters foi feita com auxílio do software Microsoft Excel, que possibilitou calcular a frequência absoluta e relativa e projetar gráficos. Os principais resultados mostram uma diferença entre cidade de porte médio, médio-grande e metrópole. As cidades médias foram caracterizadas por violência praticada por mulheres, em sua maioria autoprovocada, enquanto as cidades de porte médio-grande e metrópole são caracterizadas por agressores homens com predomínio de cônjuge e ex-cônjuge e casos que ocorreram durante datas festivas de 2017. Entre os clusters prevalece vítimas de etnia branca de 0 a 44 anos. A partir do conjunto de procedimentos realizados no processo de descoberta de conhecimento foi possível caracterizar grupos, comparar e analisar os casos de violência de gênero no Estado de São Paulo.

ABSTRACT: Brazil is the 18th country in Latin America to have a law for cases of domestic violence against women, the Maria da Penha Law. Cases of violence against women increased more than 150% in Brazil between 2016 and 2017. In this way, the use of information technologies can contribute to improve the efficiency of the management of these data. Data mining techniques make it possible to manipulate large sets of data and identify new patterns in those sets. Thus, the objective of this work was to use the experimental approach to analyze and characterize possible groups to elucidate the behavior of cases of violence against women, contributing to decision-making in actions and/or policies to combat violence against women. Initially, data were collected from SINAN/DATASUS on Domestic and Sexual Violence and other violence suffered by women. Next, pre-processing data activities were carried out; The clustering method was then applied, through the Weka tool, to facilitate the understanding of the factors that influence violence against women. Cluster analysis was performed using Microsoft Excel software, which made it possible to calculate the absolute and relative frequency and design graphs. The main results showed a difference between medium-sized, medium-large cities and metropolises. Medium-sized cities were characterized by violence perpetrated by women, mostly self-inflicted, while medium-large cities and metropolises are characterized by male aggressors with a predominance of spouse and ex-spouse and cases that occurred during festive dates in 2017. Between the clusters prevails victims of white ethnicity from 0 to 44 years old. From the set of procedures performed in the knowledge discovery process, it was possible to characterize groups, compare and analyze cases of gender violence in the State of São Paulo.

PALAVRAS-CHAVE: Violência contra a mulher. Violência doméstica. Feminicídio. Mineração de dados. *Clusterização*.

KEYWORD: Violence against women. Domestic violence. Femicide. Data mining. Clustering.

1 INTRODUÇÃO

A palavra violência é de etimologia latina, *violentia*, está relacionada a ação de profanar, agir brutalmente com força, contra a vontade do outro, causando constrangimento e noção de superioridade. Vinculado a esse vocábulo, o termo violência de gênero (também entendida como violência contra a mulher) foi uma expressão levantada pelo movimento feminista no século passado a fim de desnaturalizar e ressaltar a desigualdade de gênero, o qual envolve ameaças e agressões em todos os âmbitos e em todas as modalidades institucionais sociais (GUIMARÃES; PEDROZA, 2015).

A partir da Lei Maria da Penha, o Brasil conta com um aparato legal para casos de violência doméstica e familiar contra o sexo feminino. No art. 5ª da Lei nº 11.340, de 7 de agosto de 2006 (BRASIL, 2006), a violência contra a mulher é definida como “qualquer ação ou omissão baseada no gênero que lhe cause morte, lesão, sofrimento físico, sexual ou psicológico e dano moral ou patrimonial”.

A Sala de Apoio à Gestão Estratégica (SAGE) do Ministério da Saúde notificou que os casos de violência doméstica, sexual e outras filtradas segundo o sexo feminino aumentou em 150% entre o ano de 2016 e 2017 (SAGE, 2019). A SAGE utilizou como fonte a base de dados fornecida pelo Sistema de Informações de Agravos de Notificação (SINAN) e pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Para superar a demanda de dados gerados e acumulados em grande escala, como o caso dos registros do SINAN, surge a Descoberta do Conhecimento em Base de Dados, também conhecida pela sigla KDD (do inglês, *Knowledge Discovery in Database*). O KDD é divido em três etapas operacionais: pré-processamento, mineração de dados e pós-processamento.

Compreendendo o aumento exponencial nos números de casos de violência contra a mulher e o alcance da mineração de dados, o objetivo deste trabalho foi analisar e caracterizar possíveis agrupamentos que possam explicitar o comportamento dos dados, descrever a situação dos casos, comparar os perfis de agressão contra a mulher no Estado de São Paulo no período de 2013 a 2017 a partir dos conceitos da pesquisa de natureza experimental.

O trabalho foi desenvolvido verificando a influência de variáveis no contexto estabelecido e identificação de possíveis semelhanças para interpretação dos resultados de forma a contribuir com a tomada de decisão das ações e/ou políticas de combate à violência contra a mulher.

2 METODOLOGIA

Gil (2008) afirma que em uma pesquisa experimental se determina um objeto de estudo, seleciona-se as variáveis que seriam capazes de influenciá-lo, define-se as formas de controle e de observação dos efeitos que a variável produz no objeto. Sendo assim, este estudo trata-se de uma pesquisa de natureza experimental, em que o objeto de estudo é a agressão às mulheres no Estado de São Paulo e a análise das características dessas agressões.

Neste trabalho desenvolveu-se o processo de KDD, passando pelas etapas de pré-processamento, mineração de dados e pós-processamento.

Os dados foram extraídos a partir da plataforma do governo intitulado DATASUS, onde também foi feito o download do programa Tabwin, uma ferramenta com função de tabulação de dados (baseado nos dados do SUS) que foi utilizado para conversão de extensão e análise de frequência.

Para limpeza do *dataset* durante o pré-processamento foi utilizado o aplicativo desenvolvido pela Microsoft, o Excel.

Posteriormente, fez-se necessário o uso do Weka, pois oferece de forma prática e simples a realização de algumas tarefas de pré-processamento, além de métodos de mineração de dados.

Finalmente, o cálculo de frequência absoluta e relativa do resultado, para obter conhecimento que pudesse proporcionar a análise no pós-processamento, foi usado novamente o Excel.

2 DESENVOLVIMENTO

A mineração de dados, uma das etapas do processo do *KDD* (*Knowledge Discovery in Databases*), pós a fase de pré-processamento, envolve encaixar modelos ou determinar padrões a partir da observação de dados. A maioria dos métodos de mineração de dados são provenientes da área de *machine learning*, reconhecimento de padrões e estatística (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Os dois principais objetivos da mineração de dados são predição e descrição. No modelo preditivo, a partir de variáveis do banco de dados, é possível prever valores desconhecidos, enquanto no modelo descritivo é caracterizado as propriedades gerais (padrões interpretáveis) do banco de dados (HAN; KAMBER; PEI, 2012).

A mineração de dados pode utilizar vários métodos, tais como: Classificação, Regressão, *Clusterização*, Sumarização, Associação, Modelagem de Dependências e Detecção de Alterações e Desvios (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Neste trabalho foi utilizada a tarefa de *clusterização* que consiste em particionar os registros da base de dados em subconjuntos (ou *clusters*) de maneira que elementos presentes em um *cluster* compartilhem um conjunto de propriedades comuns e que os diferenciem dos elementos de outros *clusters*.

O *K-means* é uma técnica de *clusterização* por particionamento em que cada agrupamento ou *cluster* é representado pelo valor médio dos objetos dentro dele. Os dados são divididos (ou particionados) em *k clusters* e cada *cluster* é representado pelo objeto ou dado central, chamado de centroide (HAN; KAMBER; PEI, 2012).

A partir do estudo dos conceitos de mineração de dados e das características e comportamento da violência doméstica foi realizado um processo completo de KDD, conforme os resultados apresentados a seguir.

4 RESULTADOS OBTIDOS

Inicialmente foram extraídos os arquivos de dados classificados como “Violência doméstica, sexual e/ou outras violências” de 2013 a 2017 filtrados pela unidade federativa de São Paulo na plataforma DATASUS, por meio do qual também foi feito o *download* do arquivo de instalação do TabWin, o qual foi essencial para modificar a extensão do arquivo e realizar tabulações.

A partir disso, o arquivo passou por uma sequência de tarefas de limpeza e adequação para análise dentro do Excel: foram retiradas as colunas que não seriam relevantes para o estudo; em seguida, ao identificar campos nulos, vazios ou com asterisco foram atribuídos valores numéricos padrões conforme a legenda disponibilizada pelo DATASUS; posteriormente foram filtrados registros em que a vítima é do sexo feminino e casos que ocorreram casos entre 01/01/2013 e 31/12/2017, e por último foram removidas as colunas irrelevantes para a análise, assim como todos os registros que o atributo idade não foi preenchido ou foi ignorado, ou ainda aqueles em que o município de ocorrência foi fora do Estado de São Paulo, estava em branco ou foi ignorado. Para a verificação e garantia de integridade, essas atribuições foram baseadas em tabulações de frequência de cada atributo no TabWin.

Foi criado um atributo intitulado “MN_OCOR_C” que classifica o município da ocorrência em 10 valores categóricos diferentes, como mostra a Tabela 1, baseado na quantidade de habitantes, de acordo com o censo disponibilizado pelo IBGE (2010). A inclusão desse atributo foi realizada para possibilitar o relacionamento dos casos de agressão e a distribuição populacional por intervalo.

Tabela 1 - Frequência absoluta de cidades e registros segundo a classificação do munícipio

Classificação	Intervalo de habitantes	Frequência absoluta de cidades	Frequência absoluta de registros
1	< 5.000	159	2519
2	> 5.000 e < 10.000	122	4371
3	> 10.000 e < 20.000	122	8927
4	> 20.000 e < 50.000	119	17523
5	> 20.000 e < 100.000	49	18712
6	> 100.000 e < 250.000	49	30747
7	> 250.000 e < 500.000	17	31308
8	> 500.000 e < 1.000.000	6	23733
9	> 1.000.000 e < 5.000.000	2	10172
10	> 5.00.000 e < 15.000.000	1	31197

Fonte: Elaborado pelas autoras

Em seguida, utilizando o Weka, o arquivo resultante CSV foi convertido para a extensão ARFF, formato próprio da ferramenta. Também foram aplicados alguns filtros de pré-processamento: todos os atributos, exceto a idade, foram transformados de numéricos para nominais, uma vez que são categorias e não devem ser tratados como números; em segundo lugar, a coluna idade foi *discretizada* para que esses valores fossem colocados em faixas possibilitando um melhor controle do agrupamento de dados com a definição do valor de *bins* (número de grupos) como 5, conforme mostra o resultado na Tabela 2.

Tabela 2 **-** *Discretização* do atributo idade

Classificação	Frequência absoluta
(-inf-22]	68536
(22-45]	82235
(45-67]	24209
(67-90]	4005
(90-inf)	224

Fonte: Elaborado pelas autoras

A partir o algoritmo *k-means*, ainda dentro do Weka, foram realizadas 4 rodadas utilizando 3, 4, 5 e 6 clusters. Ao analisar os resultados obtidos, foi possível chegar à conclusão que a quantidade de *cluster* com valores de centroides mais discriminantes, foi o experimento com 5 *clusters*. Os *clusters* formados possuem a distribuição de frequência representadas na Tabela 3 - Distribuição de frequência de registros com 5 clusters.

Tabela 3 - Distribuição de frequência de registros com 5 clusters

Cluster	Registros	Porcentagem
0	24356	14%
1	50057	28%
2	40260	22%
3	37968	21%
4	26568	15%

Fonte: Elaborado pelas autoras

Para a análise dos dados foram montadas tabelas de frequência absoluta, ou seja, a quantidade de vezes que um mesmo valor do atributo foi repetido, e tabelas de frequência relativa (resultado da divisão da frequência absoluta e a quantidade de elementos). A partir dessas tabelas foi possível resumir as principais características dos *clusters* observado no Quadro *1*.

Quadro 1 - Resumo das principais características dos clusters

ATRIBUTOS	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
MN_OCOR_C	Cidades médias	Cidades médias	Cidades médias-grandes	Cidades grandes-metrópoles	Cidades médias-grandes
NU_IDADE_N	22 a 44 anos	0 a 21 anos	0 a 21 anos	22 a 44 anos	22 a 44 anos
CS_RACA	Branca	Branca	Parda	Branca	Branca
CS_ESCOL_N	Não aplicável	Não aplicável	Não aplicável	Não aplicável	Ensino médio completo
DT_OCOR	12/10/2017	20/04/2016	01/01/2017	01/01/2017	19/02/2017
AUTOR_SEXO	Feminino	Feminino	Masculino	Masculino	Masculino
EVOLUCAO	Ign/Branco	Ign/Branco	Ign/Branco	Ign/Branco	Ign/Branco
LES_AUTOP	Sim	Não	Não	Não	Não
Tipo de Violência	Outras violências	Agressão física	Agressão física	Agressão física	Agressão física
Vínculo com o agressor	Própria pessoa	Mãe, conhecido	Cônjuge, ex-cônjuge, conhecido	Cônjuge	Cônjuge, ex-cônjuge

Fonte: Elaborado pelas autoras

Com a análise dos *clusters* pode-se observar que as agressões físicas são autoprovocadas ou causadas por agressores do sexo masculino, na maioria das vezes pelo cônjuge e pelo ex-cônjuge, representados nos *cluster* 2 e 3.

Os casos de cidades médias concentram casos ocasionados por mulheres, ou seja, autoprovocadas e pela mãe, como mostra os *cluster* 0 e 1.

As ocorrências de cidades médias a metrópoles, *clusters* 2, 3 e 4, envolvem agressores masculinos e datam o ano de 2017, sendo os *clusters* 2 e 3 com alta frequência de registros no Ano Novo de 2017 (01 de janeiro) e o *cluster* 4 no pré-carnaval de 2017 (19 de fevereiro) o que pode ser justificado pelo aumento de consumo de álcool e drogas nessas épocas festivas.

Quanto a etnia da vítima observa-se o fenômeno da maioria das vítimas serem brancas, embora, no Brasil, a maior parte da população seja preta. É possível levantar a hipótese de que mulheres brancas confiam mais na polícia e tem mais acessibilidade dos que as pardas e pretas.

A escolaridade tem como maiores frequências casos que não tiveram o atributo preenchido, ou foi ignorado, provavelmente porque tem maior frequência de vítimas de 0 a 21 anos, que engloba crianças ainda sem a escolaridade definida. Somente no *cluster* 4 é observado o ensino médio completo de forma acentuada e consequentemente com vítimas acima de 22 anos.

5 CONSIDERAÇÕES FINAIS

Por meio dos resultados obtidos e análises realizadas, entende-se que este trabalho conseguiu atingir o objetivo proposto expondo as principais características, descrevendo a situação dos casos, comparando perfis de agressão e possibilitando a contribuição para a tomada de decisão nas ações e/ou políticas públicas de combate à violência doméstica, sexual e outras contra a mulher no Estado de São Paulo.

REFERÊNCIAS

BRASIL. Lei nº 11.340, de 7 de agosto de 2006. Cria mecanismos para coibir a violência doméstica e familiar contra a mulher, nos termos do § 8º do art. 226 da Constituição Federal, da Convenção sobre a Eliminação de Todas as Formas de Discriminação contra as Mulheres e da Convenção Interamericana para Prevenir, Punir e Erradicar a Violência contra a Mulher; dispõe sobre a criação dos Juizados de Violência Doméstica e Familiar contra a Mulher; altera o Código de Processo Penal, o Código Penal e a Lei de Execução Penal; e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm>. Acesso em: 27 jul. 2021.

CARAVANTES, L. Violência intrafamiliar en la reforma del sector salud. In: COSTA, A.M.; MERCHÁN-HAMANN, E.; TAJER, D. (Orgs.). **Saúde, eqüidade e gênero: um desafio para as políticas públicas**. Brasília: Editora Universidade de Brasília, 2000. p.18.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37, 15 Mar. 1996.

GIL, Antonio Carlos. **Métodos e Técnicas de Pesquisa Social**. São Paulo: Atlas, 2008.

GUIMARÃES, Maisa Campos; PEDROZA, Regina Lucia Sucupira. VIOLÊNCIA CONTRA A MULHER: problematizando definições teóricas, filosóficas e jurídicas. Psicologia & Sociedade, [S.L.], v. 27, n. 2, p. 256-266, ago. 2015. FapUNIFESP (SciELO). <http://dx.doi.org/10.1590/1807-03102015v27n2p256>.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. São Francisco, EUA: Morgan Kaufmann Publishers, 2011.

IBGE. **Cidades@**. 2010. Disponível em: <https://cidades.ibge.gov.br/brasil/sintese/sp?indicadores=25207>. Acesso em: 27 jul. 2021.

SAGE. Ministério da Saúde. **Violências Domésticas, Sexual e Outras**. 2019. Disponível em: <https://portalsage.saude.gov.br/painelManutencao/Viol%C3%AAncias%20Dom%C3%A9sticas,%20Sexual%20e%20Outras>. Acesso em: 25 out. 2020.