

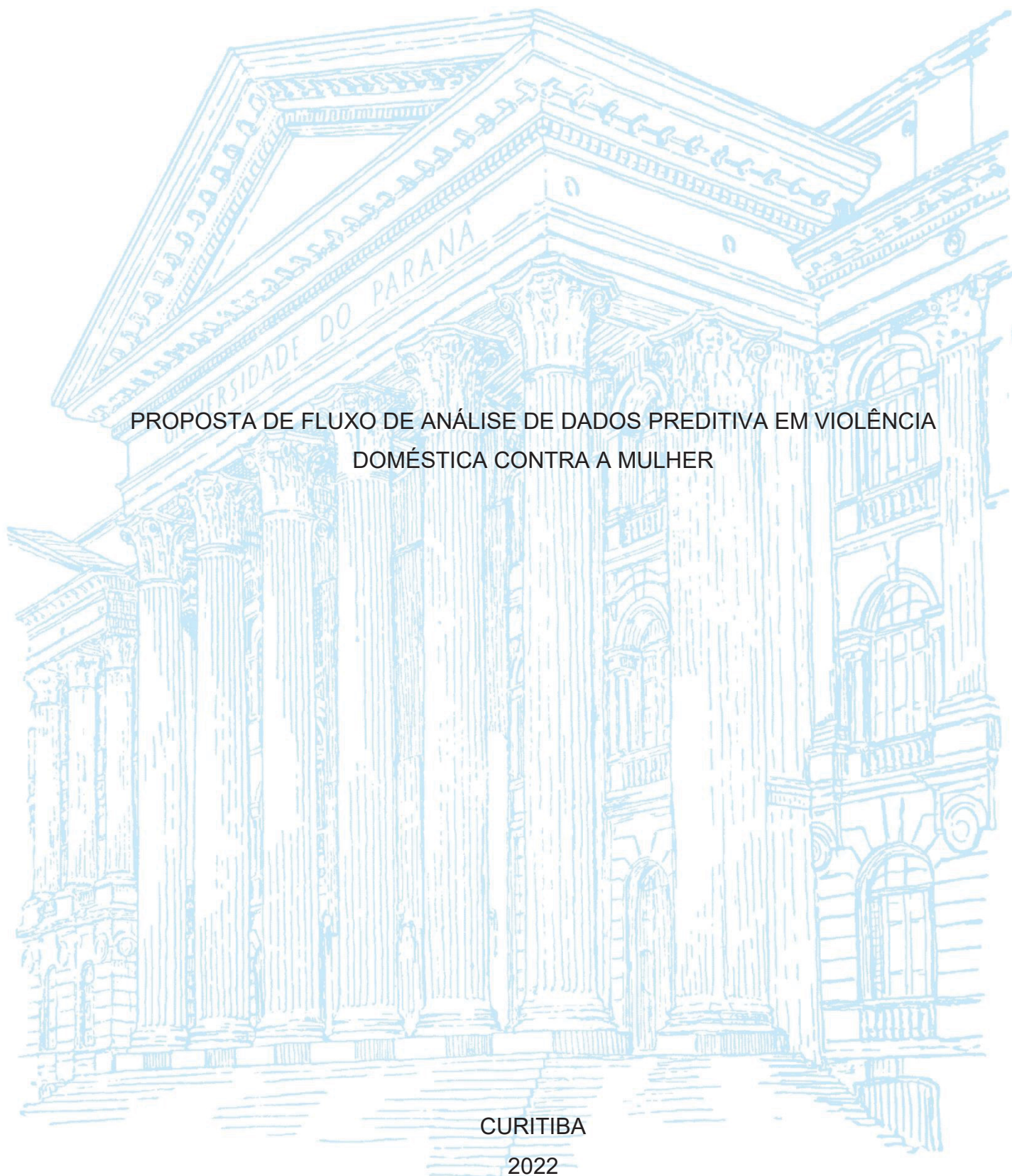
UNIVERSIDADE FEDERAL DO PARANÁ

CHRISTIAN ARISTÓTELES DA SILVA COSTA

PROPOSTA DE FLUXO DE ANÁLISE DE DADOS PREDITIVA EM VIOLÊNCIA
DOMÉSTICA CONTRA A MULHER

CURITIBA

2022



CHRISTIAN ARISTÓTELES DA SILVA COSTA

PROPOSTA DE FLUXO DE ANÁLISE DE DADOS PREDITIVA EM VIOLÊNCIA
DOMÉSTICA CONTRA A MULHER

Dissertação apresentada ao curso de Pós-Graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Gestão da Informação.

Orientador(a): Prof^ª. Dr^a. Denise Fukumi Tsunoda.

CURITIBA

2022

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS

Costa, Christian Aristóteles da Silva

Proposta de fluxo de análise de dados preditiva em violência doméstica contra a mulher / Christian Aristóteles da Silva Costa. – Curitiba, 2022.

1 recurso on-line : PDF.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Gestão da Informação.

Orientadora: Profa. Dra. Denise Fukumi Tsunoda.

1. Gestão da Informação. 2. Violência doméstica.
3. Mineração de dados. I. Tsunoda, Denise Fukumi.
II. Universidade Federal do Paraná. Programa de Pós-Graduação em Gestão da Informação. III. Título.

Bibliotecária: Maria Lidiane Herculano Graciosa CRB-9/2008



MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA
INFORMAÇÃO - 40001016058P1

TERMO DE APROVAÇÃO

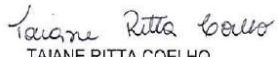
Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **CHRISTIAN ARISTÓTELES DA SILVA COSTA** intitulada: **PROPOSTA DE FLUXO DE ANÁLISE DE DADOS PREDITIVA EM VIOLÊNCIA DOMÉSTICA CONTRA A MULHER**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 12 de Setembro de 2022.


JOSE MARCELO ALMEIDA PRADO CESTARI
Presidente da Banca Examinadora


LUIZ PANTHOCA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)


TAIANE RITTA COELHO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

AGRADECIMENTOS

Agradeço a todos que contribuíram para esta dissertação em especial a minha orientadora Prof.^a Dr.^a Denise F. Tsunoda que contribuiu com seu tempo e experiência.

Agradeço à Secretaria de Segurança Pública (SESP) e ao Centro de Análise, Planejamento e Estatística (CAPE) pelo termo de cooperação técnica e disponibilização dos dados utilizados nas análises.

Agradeço à Secretaria do Programa Pós-graduação em Gestão da Informação (PPGGI) por todo auxílio e atenção dispensados.

RESUMO

A pesquisa trata da violência doméstica contra mulher com motivação acadêmica e social sob a visão da análise de dados. O estudo propõe a criação de um fluxo de análise de dados preditiva para vítimas fatais de violência doméstica utilizando métodos de “Knowledge Discovery in Databases” (KDD). A etapa da mineração de dados foi aplicada em duas bases de dados: a do Boletim de Ocorrência Unificado e a do Sistema de Ocorrências Letais, ambas da Secretaria de Segurança Pública do Paraná. Essas bases contêm dados de mulheres vítimas de violência doméstica e feminicídio dos anos de 2018, 2019 e 2020. Por meio das análises realizadas foi possível criar um perfil para vítimas de violência doméstica e vítimas fatais de feminicídio utilizando as frequências dos atributos com maior capacidade de explicação. As vítimas fatais de violência contra mulher, feminicídio da base de dados de Ocorrências Letais tem em média de 37 anos. Das mulheres mortas por esse crime nos três anos, 76,7% morreram por arma branca ou imprópria, por arma de fogo ou por agressão física. Pouco mais de 57% dessas vítimas foram declaradas brancas e 28% declaradas pardas. Mais de 65% desses crimes ocorreram em 14 municípios do estado do Paraná. A análise com o KDD foi aplicada às bases de ocorrências letais e naturezas jurídicas por vítima. Utilizando técnicas de seleção de atributos (CFS, Info Gain, Gain Ratio e ReliefF) foi possível determinar os atributos mais relevantes para a base de dados de naturezas jurídicas. Os atributos mais relevantes para cada técnica foram submetidos aos algoritmos de indução de regras PRISM e CN2. Os resultados do PRISM apresentam acurácia abaixo de 0,548 e no máximo 16% de acertos na predição da classe “S”. Os resultados do CN2 usando os atributos mais relevantes segundo as técnicas CFS e Info Gain, apresentam acurácia acima de 0,71 e proporção de acertos para duas classes mais equilibrados que as outras técnicas. O resultado da análise preditiva foi aprovado por dois especialistas do Centro de Análise Planejamento e Estatística e oportunizou a criação de três fluxos de análise de dados que podem ser aplicados em outras bases de dados de violência doméstica contra mulher.

Palavras-chave: KDD. Violência Doméstica. Mineração de Dados. Preditivo. Gestão da Informação.

ABSTRACT

The research deals with domestic violence against women with academic and social motivation from the point of view of data analysis. The study proposes the creation of a predictive data analysis flow for fatal victims of domestic violence using “Knowledge Discovery in Databases” (KDD) methods. The data mining step was applied in two databases: the Boletim de Ocorrência Unificado and the Lethal Occurrence System, both from the Paraná Public Security Department. These databases contain data on women victims of domestic violence and femicide from the years 2018, 2019 and 2020. Through the analyzes carried out, it was possible to create a profile for victims of domestic violence and fatal victims of femicide using the frequencies of the attributes with the greatest ability to explanation. The fatal victims of violence against women, femicide in the Lethal Occurrences database have an average of 37 years. Of the women killed by this crime in the three years, 76.7% died from a bladed weapon or improper weapon, fire gun or physical aggression. Just over 57% of these victims were declared white and 28% declared brown. More than 65% of these crimes occurred in 14 municipalities in the state of Paraná. The analysis with the KDD was applied to the bases of lethal occurrences and legal nature per victim. Using attribute selection techniques (CFS, Info Gain, Gain Ratio and ReliefF) it was possible to determine the most relevant attributes for the legal database. The most relevant attributes for each technique were submitted to the PRISM and CN2 rule induction algorithms. The PRISM results present accuracy below 0.548 and a maximum of 16% of hits in the prediction of the “S” class. The CN2 results using the most relevant attributes according to the CFS and Info Gain techniques, present accuracy above 0.71 and the proportion of hits for two classes more balanced than the other techniques. The result of the predictive analysis was approved by two specialists from the Center for Analysis, Planning and Statistics and provided the opportunity to create three data analysis flows that can be applied to other databases on domestic violence against women.

Keywords: KDD. Domestic violence. Data Mining. Predictive. Information management.

LISTA DE FIGURAS

FIGURA 1 – ETAPAS DO KDD	30
FIGURA 2 – COMPARAÇÃO ENTRE KDD, SEMMA E CRISP-DM	31
FIGURA 3 - MATRIZES DE CONFUSÃO GERADAS PELO ALGORITMO CN2	97
FIGURA 4 - FLUXOGRAMA DE ANÁLISE DE DADOS PREDITIVA COM ESTATÍSTICA DESCRITIVA.....	100
FIGURA 5 - FLUXOGRAMA DE ANÁLISE DE DADOS PREDITIVA COM KDD	103
FIGURA 6 - FLUXOGRAMA DE ANÁLISE DE DADOS PREDITIVA COM KDD PARA BASES COM REINCIDÊNCIAS	105

LISTA DE GRÁFICOS

GRÁFICO 1 – COMPARATIVO DAS IDADES DAS VÍTIMAS DE FEMINICÍDIO DA BASE SCOL AGRUPADA POR ANO	66
GRÁFICO 2 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DOS AUTORES CONHECIDOS EM FEMINICÍDIOS DA BASE SCOL AGRUPADOS POR ANO.....	67
GRÁFICO 3 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DOS MEIOS EMPREGADOS NOS FEMINICÍDIOS DA BASE SCOL AGRUPADOS POR ANO.....	68
GRÁFICO 4 – COMPARATIVO DA FREQUÊNCIA RELATIVA DA COR DA PELE DAS VÍTIMAS DE FEMINICÍDIO DO SCOL	69
GRÁFICO 5 – COMPARATIVO DE FREQUÊNCIA ABSOLUTA DAS PROFISSÕES DAS VÍTIMAS DE FEMINICÍDIO NO SCOL	70
GRÁFICO 6 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO ESTADO CIVIL DAS VÍTIMAS DE FEMINICÍDIO DO SCOL AGRUPADOS POR ANO	71
GRÁFICO 7 - COMPARATIVO DA FREQUÊNCIA RELATIVA DE FEMINICÍDIOS POR DIA DA SEMANA AGRUPADOS POR ANO	72
GRÁFICO 8 - COMPARATIVO DE IDADES ESTIMADAS DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA NO BOU AGRUPADAS POR ANO	75
GRÁFICO 9 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DA COR DA PELE DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADAS POR ANO.....	76
GRÁFICO 10 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO ESTADO CIVIL DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADOS POR ANO.....	77
GRÁFICO 11 – COMPARATIVO DA FREQUÊNCIA RELATIVA DO GRAU DE INSTRUÇÃO DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADOS POR ANO.....	78
GRÁFICO 12 - COMPARATIVO DE FREQUÊNCIA RELATIVA DE BOLETINS DO BOU POR DIA DA SEMANA AGRUPADOS POR ANO	79

GRÁFICO 13 - COMPARATIVO DA FREQUÊNCIA ABSOLUTA DA OCUPAÇÃO DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADAS POR ANO.....	81
GRÁFICO 14 - COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO MEIO EMPREGADO NAS OCORRÊNCIAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADOS POR ANO.....	82
GRÁFICO 15 - COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO NOME JURÍDICO EM OCORRÊNCIAS DE VIOLÊNCIA DOMÉSTICA AGRUPADOS POR ANO.....	84

LISTA DE QUADROS

QUADRO 1 – TIPOS DE VIOLÊNCIA DOMÉSTICA E FAMILIAR CONTRA A MULHER.....	25
QUADRO 2 – DADOS, INFORMAÇÃO E CONHECIMENTO	27
QUADRO 3 – ABORDAGENS DE SELEÇÃO DE ATRIBUTOS	35
QUADRO 4 - MATRIZ DE CONFUSÃO DE CLASSE BINÁRIA.....	40
QUADRO 5 - MEDIDAS PARA AVALIAÇÃO DE MODELO PREDITIVO COM CLASSIFICADORES	41
QUADRO 6 – TERMOS DE BUSCA EM PORTUGUÊS E INGLÊS APLICADOS NAS BASES SCOPUS E WEB OF SCIENCE.....	43
QUADRO 7 – CLASSIFICAÇÃO DA PESQUISA.....	50
QUADRO 8 – EXEMPLO DE LINHA COM DADOS AGRUPADOS POR VÍTIMA	53
QUADRO 9 - SIMPLIFICAÇÃO DE VALORES NO ATRIBUTO NATUREZA JURÍDICA DO BOU	55
QUADRO 10 – MODIFICAÇÃO DOS VALORES DO ATRIBUTO DIASEMANA DO BOU	56
QUADRO 11 - MODIFICAÇÃO DOS VALORES DO ATRIBUTO PERÍODO DO BOU	56
QUADRO 12 - MODIFICAÇÃO DOS VALORES DO ATRIBUTO DIA DA SEMANA DO SCOL	56
QUADRO 13 – ATRIBUTOS DA BASE SCOL	59
QUADRO 14 – ATRIBUTOS DA BASE BOU	62
QUADRO 15 – ATRIBUTOS DA BASE NATUREZAS	63
QUADRO 16 - COMPARATIVO DA ANÁLISE ESTATÍSTICA DAS BASES BOU E SCOL	88
QUADRO 17 - FREQUÊNCIA DAS NATUREZAS JURÍDICAS NAS OCORRÊNCIAS DO BOU AGRUPADAS POR VÍTIMA DE FEMINICÍDIO	90
QUADRO 18 - ATRIBUTOS SELECIONADOS NA BASE DE NATUREZAS JURÍDICAS DAS VÍTIMAS DE FEMINICÍDIO E VIOLÊNCIA DOMÉSTICA.....	95
QUADRO 19 - REGRAS INDUZIDAS COM AS TÉCNICAS DE SELEÇÃO DE ATRIBUTOS CFS E INFO GAIN.....	98

LISTA DE TABELAS

TABELA 1 - RETORNO DOS TERMOS DE BUSCA NO PORTAL CAPES.....	19
TABELA 2 - ESTATÍSTICA DESCRITIVA DO ATRIBUTO IDADE DA BASE SCOL 66	
TABELA 3 – ESTATÍSTICA DESCRITIVA DAS IDADES ESTIMADAS REGISTRADA NO BOU AGRUPADAS POR ANO	74
TABELA 4 - RESULTADOS DA EXECUÇÃO DO ALGORITMOMO CN2 COM AS TÉCNICAS DE SELEÇÃO	96

LISTA DE ABREVIATURAS OU SIGLAS

BOU	- Boletim de Ocorrência Unificado
CAPE	- Centro de Análise, Planejamento e Estatística
CELEPAR	- Companhia de Tecnologia da Informação e Comunicação do Paraná
CSV	- Comma Separated Values
KDD	- Descoberta de Conhecimento em Base de Dados
FCA	- Formal Concept Analysis
GI	- Gestão da Informação
LGPD	- Lei Geral de Proteção de Dados Pessoais
MD	- Mineração de Dados
PCPR	- Polícia Civil do Paraná
PMPR	- Polícia Militar do Paraná
PPGGI	- Programa de Pós-Graduação em Gestão da Informação
RNA	- Redes Neurais Artificiais
SCOL	- Sistema de Ocorrências Letais
SENASP	- Secretaria Nacional de Segurança Pública
SESP/PR	- Secretaria de Segurança Pública do Estado do Paraná
SVM	- Support Vector Machine
VD	- Violência Doméstica
WoS	- Web of Science

SUMÁRIO

1 INTRODUÇÃO	15
1.1 PROBLEMA	16
1.2 OBJETIVOS	18
1.2.1 Objetivo geral	18
1.2.2 Objetivos específicos.....	18
1.3 JUSTIFICATIVA	18
1.3.1 Justificativa acadêmica.....	19
1.3.2 Justificativa ao Programa de Pós-Graduação em Gestão da Informação	20
1.3.3 Justificativa à sociedade.....	20
1.4 DELIMITAÇÕES DA PESQUISA	21
1.5 ESTRUTURA DO DOCUMENTO.....	21
2 REVISÃO DE LITERATURA	23
2.1 VIOLÊNCIA DOMÉSTICA	23
2.2 GESTÃO DA INFORMAÇÃO	27
2.3 RECUPERAÇÃO DA INFORMAÇÃO	28
2.4 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (KDD)	29
2.4.1 Data Mining	31
2.4.2 Tarefas de mineração de dados	32
2.5 REAMOSTRAGEM (RESAMPLING).....	33
2.5.1 Oversampling	33
2.5.2 Undersampling	34
2.6 SELEÇÃO DE ATRIBUTOS	34
2.6.1 Correlation-based feature selection (CFS)	36
2.6.2 Info Gain.....	36
2.6.3 Gain Ratio	37
2.6.4 ReliefF	37
2.7 REGRAS DE CLASSIFICAÇÃO.....	37
2.7.1 Algoritmo PRISM.....	38
2.7.2 Algoritmo CN2	39
2.8 AVALIAÇÃO DE MODELOS PREDITIVOS DE CLASSIFICADORES	39
2.8.1 Validação cruzada	41
2.9 ESTATÍSTICA DESCRITIVA.....	42

2.10 VIOLÊNCIA DOMÉSTICA E ANÁLISE DE DADOS	43
2.10.1 Métodos numéricos utilizados em pesquisas de VD	45
2.11 SOFTWARES.....	45
2.11.1 Software Orange	46
2.11.2 Software RStudio	47
2.11.3 Software WEKA.....	47
3 ENCAMINHAMENTOS METODOLÓGICOS	49
3.1 CARACTERIZAÇÃO DE PESQUISA	49
3.2 MATERIAIS E MÉTODOS.....	50
3.2.1 Modelo atual de recuperação de dados e produção de relatórios no CAPE	50
3.2.2 Seleção de dados.....	52
3.2.3 Pré-processamento de dados	52
3.2.4 Transformação dos dados.....	55
3.2.5 Mineração de dados	57
3.3 TABELAS DA PESQUISA	57
3.3.1 Base SCOL	57
3.3.2 Base BOU	61
3.3.3 Base Naturezas	63
3.4 AVALIAÇÃO DO MODELO PREDITIVO	64
4 RESULTADOS E ANÁLISES	65
4.1 RESULTADOS DA ESTATÍSTICA DESCRITIVA.....	65
4.1.1 Estatística descritiva da base SCOL	65
4.1.2 Estatística descritiva da base BOU	74
4.1.3 Estatística descritiva da base SCOL <i>versus</i> BOU	87
4.1.4 Estatística descritiva da Base Naturezas <i>versus</i> feminicídios	89
4.2 RESULTADOS DA MINERAÇÃO DE DADOS COM ALGORITMOS PREDITIVOS DE CLASSIFICAÇÃO	94
4.2.1 Seleção de atributos das bases	94
4.2.2 Resultados dos algoritmos CN2 e PRISM.....	95
4.3 FLUXO DE ANÁLISE PREDITIVA DE DADOS USANDO KDD EM DADOS DE VÍTIMAS FATAIS	99
5 CONSIDERAÇÕES FINAIS	106
REFERÊNCIAS.....	109

1 INTRODUÇÃO

A análise de dados de violência doméstica (VD) e feminicídio desta dissertação, expõe alguns aspectos negativos da sociedade paranaense quanto a violência contra à mulher.

Max Horkheimer ainda na década de 1930 escrevia sobre emancipação social e como consequências dessa emancipação a liberdade e a igualdade (MELO, 2011; NOBRE, 2004). A existência da VD contra à mulher nos dados da segurança pública pode ser considerado um obstáculo para esta emancipação idealizada por autores da teoria crítica como Max Horkheimer, já que a VD praticada contra mulher ou priva sua vítima da liberdade e/ou expõe a desigualdade com que a mulher é tratada nas suas relações cotidianas.

A análise crítica, diferentemente das outras teorias, não se limita a mera descrição esquecendo da dimensão temporal, mas sim, pretende criar um diagnóstico temporal, examinado o existente, identificando e analisando obstáculos à emancipação (MELO, 2011; NOBRE, 2004). Portanto, estudar a VD academicamente contribui de forma consistente para criação de diagnósticos territoriais e temporais sobre esse tema.

Para uma visão inicial desse trabalho são descritos conceitos sobre o objeto da pesquisa, a gestão da informação, a descoberta de conhecimento em base de dados (KDD), técnicas de amostragem, técnicas de seleção de atributos, regras de classificação, avaliação de métodos preditivos, softwares utilizados na análise e a análise de dados sobre violência doméstica (revisão integrativa). Esses conceitos, além de descrever as ideias principais que norteiam a pesquisa, foram pensados também para demonstrar caminhos científicos (como a KDD) para criar conhecimento novo sobre o objeto da pesquisa.

Até que se sabe, uma pesquisa conduzida com o KDD em dados de VD da segurança pública do Paraná (SESP/PR), sob o mesmo prisma, é a primeira tanto para segurança pública do estado quanto para UFPR, provavelmente pela sensibilidade dos dados, o que dificulta seu acesso pela proteção legal. A Lei Geral de Proteção de Dados Pessoais (LGPD), Lei Federal nº 13.709, define dados sensíveis como:

Dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural. (BRASIL, 2018).

As bases de dados da SESP/PR armazenam diversos dados sensíveis, o que torna o seu direito de uso vedado a própria administração pública.

Para fins desta pesquisa a SESP/PR autorizou a extração de alguns atributos do banco de dados do boletim de ocorrência unificado (BOU) e do banco de dados do sistema de ocorrências letais (SCOL). Dessa extração resultaram, uma base de dados formada de atributos do BOU relacionados a ocorrências de VD, uma base de dados somente de vítimas de feminicídio e outra base de dados formada pela contagem de naturezas jurídicas das ocorrências agrupadas por vítima.

A análise sobre essas bases de dados oportunizou a criação de um fluxo de análise preditiva para dados de VD oriundos de ocorrências criminais e em extrair padrões utilizando o KDD e a estatística descritiva.

Esse estudo acrescenta novos conhecimentos sobre o uso de KDD aplicado em bases de dados sobre VD contra mulher. Seus resultados podem incentivar novas pesquisas e apoiar as tomadas de decisões dos gestores de segurança pública.

1.1 PROBLEMA

A gestão dos sistemas de segurança pública, o tratamento e o monitoramento dos dados de atividades criminosas são fundamentais para tomada de decisão. Um exemplo, a alocação de efetivo policial e reforço são decisões dos gestores locais e a SESP/PR tem gestão sobre eles, portanto, é mais comum encontrar reforços policiais em locais de muito trânsito de pessoas ou onde as ocorrências criminais estão mais agrupadas.

Essa é uma decisão que não se baseia exclusivamente em dados, existem outras variáveis envolvidas nessa decisão, muitas vezes envolvem a geografia, o conhecimento do efetivo, o conhecimento sobre quais veículos e equipamentos disponíveis, além de conhecimento sobre o tráfego de pessoas na região no decorrer do tempo.

Porém, sem os dados de quantidade de efetivo disponível, dados das ocorrências, dados de disponibilidade de veículos e equipamentos, a decisão tomada pelos gestores pode ser equivocada, sendo ineficaz contra os problemas dos locais

que necessitam de maior policiamento. Sendo assim, a tomada de decisões da segurança pública depende de uma gestão atenta e contínua dos dados e informações para ações mais assertivas.

O tratamento de informação para monitorar às ocorrências violentas, como as de violência doméstica não possuem um fluxo padrão de trabalho ou modelo com capacidade de replicabilidade na previsão de eventos baseado em fatos passados. Atualmente, as análises sobre diversos tipos de crimes, incluindo os violentos, são realizados por analistas criminais do Centro de Análise Planejamento e Estatística (CAPE), o conhecimento de cada indivíduo sobre determinado objeto de análise é determinante no resultado do estudo, assim, uma análise feita sobre o mesmo tema por outro analista pode ter conclusões um pouco diferentes.

Ocorreu no CAPE um estudo sobre mulheres vítimas de feminicídio no estado no ano de 2017. No estado, os delegados da polícia civil entre muitas incumbências são responsáveis por classificar um crime, como o de morte por feminicídio. O analista de mortes do CAPE, sabendo que a responsabilidade de classificação de uma morte como feminicídio é dos delegados, extraiu os casos de vítimas de feminicídio registrados pela polícia civil, então, estudou os boletins registrados no BOU anteriores a morte de cada vítima.

Este analista relatou que no estudo para a maioria das vítimas existiam vários boletins registrados no BOU e o padrão visto nessas vítimas foi a existência de um tempo longo entre o último boletim registrado e o boletim registrando a morte dessa vítima. O estudo não foi repetido em outros anos.

Entre os muitos dados registrados nos sistemas da SESP/PR, alguns poderiam ser mais estudados, principalmente os envolvendo violência. Um método de análise com natureza preditiva poderia ser repetido nos estudos, este possibilitaria a comparação de seus resultados ao longo do tempo.

O conhecimento acadêmico tem muito a oferecer para diversas áreas, inclusive para administração pública. A repetição de eventos num banco de dados como o BOU, se forem corretamente tratados por fluxos padronizados e/ou modelos científicos como o KDD, tendem a acrescentar conhecimento novo e útil para a gestão. Um método, que pudesse prever a possibilidade de morte de uma vítima do BOU, poderia oportunizar a notificação de um perfil dessas vítimas às delegacias ou batalhões de área para que observassem com maior proximidade essas características.

Entretanto, para construir um perfil com base em KDD é necessário valorar os conjuntos de dados e seus atributos. Como determinar quais atributos devem ser considerados na proposta de um fluxo de análise preditiva para identificação vítimas fatais de violência doméstica?

1.2 OBJETIVOS

Este tópico está dividido em: objetivo geral e específicos.

1.2.1 Objetivo geral

Propor um fluxo de análise de dados preditiva que possibilite determinar um perfil para as mulheres vítimas fatais de violência doméstica.

1.2.2 Objetivos específicos

Derivados do objetivo geral, estão propostos os seguintes objetivos específicos:

- a) analisar e descrever o processo atual de recuperação de informação e produção de relatórios sobre o tema Violência Doméstica no Centro de Análise, Planejamento e Estatística;
- b) aplicar a Descoberta de Conhecimento em Bases de Dados para encontrar padrões nos dados de violência doméstica quando a vítima foi morta;
- c) analisar e comparar os resultados dos métodos de Indução de regras.

1.3 JUSTIFICATIVA

Este tópico foi dividido em: justificativa acadêmica, justificativa ao Programa ao Pós-graduação em Gestão da Informação e justificativa à sociedade.

1.3.1 Justificativa acadêmica

Em pesquisa no portal CAPES em fevereiro e março de 2021, utilizando o termo de pesquisa “Descoberta de Conhecimento em Base de Dados”, resultaram 11 artigos, contudo 8 com texto completo e um duplicado.

Utilizando o termo “Violência Doméstica”, resultaram 3.227 artigos, 2.909 revisados por pares e 1.989 dos últimos dez anos. Os artigos que tratavam de KDD utilizaram métodos diferentes. Métodos como APRIORI, Redes Neurais Artificiais (RNA), Algoritmo de Bayes, Mapas Auto-organizáveis (SOM), Classificação e Regressão Logística (RL) foram utilizados no processo de descoberta de conhecimento.

Foi observado o uso ferramentas livres como o R e WEKA em alguns desses artigos. Os bancos de dados estudados eram em sua maioria abertos com exceção de dois destes.

Buscando pelos termos KDD e VD nesse portal resultam artigos, porém com os mesmos termos em inglês nota-se uma produção científica absolutamente maior. Em KDD a produção em português representa menos de 4%, em VD representa 1%.

TABELA 1 - RETORNO DOS TERMOS DE BUSCA NO PORTAL CAPES

Termos de Busca com Aspas Duplas	Knowledge Discovery in Databases (nº)	Domestic Violence (nº)	Descoberta de Conhecimento em Bases de Dados (nº)	Violência Doméstica (nº)
Total	696	128.127	11	3.469
Artigos	606	86.683	10	3.227
Revisão por pares	505	72.037	10	2.909
Últimos 10 anos	237	52.172	10	1.989
Idioma Inglês	216	51.441	9	1.385
Idioma Espanhol	33	1.145	1	1.079
Idioma Português	21	902	6	947
Idioma Francês	2	463	-	11
Idioma Chinês	2	232	-	7

FONTE: O Autor (2021).

A produção científica relacionando os termos (“violência doméstica” e “descoberta de conhecimento em base de dados”) ou (“mineração de dados” ou “data mining” ou “Knowledge Discovery in Databases”) e (“Domestic Violence”) no portal CAPES lista 3 artigos utilizando “*text mining*” com alguma relação com esta pesquisa, mais 3 da base *Web of Science* (WoS), ambos publicados em inglês escritos por pesquisadores relacionados a universidades de outros países. No Brasil temos

poucos artigos listados pelo portal da CAPES no contexto dessa pesquisa e nenhum tem aderência perfeita, uma possível causa seja a sensibilidade dos dados sobre VD que são de responsabilidade do Estado no Brasil.

Em novembro de 2020 a UFPR firmou um termo de cooperação técnica com a SESP/PR, neste existe a previsão de uso de dados do estado em pesquisas dessa universidade, fato que sugere uma oportunidade que pode contribuir para novos trabalhos sobre KDD como este que estuda dados sobre VD.

1.3.2 Justificativa ao Programa de Pós-Graduação em Gestão da Informação

Em busca por dissertações e teses no sistema de bibliotecas da UFPR usando os termos (“violência doméstica” e “data mining”), como também, (“violência doméstica” e “descoberta de conhecimento em base de dados”), não foram encontradas teses ou dissertações com os referidos temas associados. Usando o termo “violência doméstica”, tem-se somente 10 dissertações e 4 teses, já com o termo “Descoberta de Conhecimento em Base de Dados” são 22 dissertações e 10 Teses. Nessas pesquisas de KDD nenhum estudo utilizou dados de VD.

Uma dissertação na área de KDD estudando dados de VD da SESP/PR é a primeira realizada no PPGGI, essa deve oferecer uma referência que pode incentivar novas pesquisas nessas áreas.

1.3.3 Justificativa à sociedade

As sociedades mais primitivas são relatadas na história oficial como violentas. A sua busca por acesso à água, alimentos e reprodução as levava a tomar posse do que os grupos mais fracos tinham a sua disposição.

As sociedades, nos últimos séculos, mudaram muito em comparação àquelas primitivas, muito se devendo aos avanços tecnológicos e científicos que proporcionaram a computação, o novo modelo de comunicação gerando novas necessidades. Mesmo tendo evoluído em muitos aspectos um problema ainda ocorre, a violência intramuros.

A legislação avançou nos últimos 16 anos criando mecanismos para a proteção da mulher, como a Lei Maria da Penha, maiores explicações são vistas no referencial teórico. Contudo, mesmo que se avance no aspecto da segurança legal

para coibir a VD contra à mulher, esta segurança nem sempre se materializa na realidade de muitas mulheres, inclusive no estado do Paraná. Fato este que se confirma ao observar a quantidade de ocorrências de violência doméstica registradas no estado.

As ocorrências que reportam VD no estado do Paraná são numerosas, 70.940 boletins em média nos 3 anos estudados (2018, 2019 e 2020), o suficiente para que um analista experiente leve muito tempo para encontrar padrões úteis que possam orientar gestores da segurança pública. Somente no ano de 2020 foram consumadas 58.816 ameaças, 39.939 Lesões Corporais decorrentes de VD, 370 lesões graves ou gravíssimas e 6 Lesões Corporais seguidas de morte contra vulnerável. Dados do BOU do estado do Paraná.

Descobrir padrões ainda desconhecidos nos dados, pode acrescentar conhecimento que auxiliará os gestores no processo de priorização no atendimento e prevenção a estes tipos de crime. Usar o conhecimento científico da área de KDD para auxiliar no combate à VD é uma oportunidade que pode evitar o sofrimento ou mal maior de suas vítimas.

1.4 DELIMITAÇÕES DA PESQUISA

Esta pesquisa se limitou a estudar violência doméstica exclusivamente contra à mulher.

Quanto aos dados estudados, a pesquisa é limitada aos atributos autorizados pelo CAPE. Esses atributos autorizados compõem três bases de dados e alguns deles devem ser anonimizados antes do depósito desses pós pesquisa, como previsto no termo de cooperação técnica.

Da VD no estado do Paraná, foi delimitado o estudo em dados de vítimas mulheres com pelo menos 18 anos que tenham pelo menos um registro de ocorrência no sistema BOU e SCOL nos anos de 2018, 2019 e 2020.

1.5 ESTRUTURA DO DOCUMENTO

A primeira seção é composta por introdução, problema, justificativas acadêmica, ao Programa de Pós-Graduação em Gestão da Informação e à sociedade, objetivos geral e específicos e delimitações da pesquisa.

A segunda seção é composta pela revisão de literatura de violência doméstica, gestão da informação, descoberta de conhecimento em bases de dados, reamostragem, seleção de atributos, regras de classificação, avaliação de modelos preditivos de classificadores, estatística descritiva, softwares, violência doméstica e análise de dados.

A terceira seção é composta pelos encaminhamentos metodológicos, que compreendem: a caracterização da pesquisa, materiais e métodos, tabelas da pesquisa e avaliação do modelo preditivo.

A quarta seção é composta pelos resultados da estatística descritiva, resultados da mineração de dados com algoritmos preditivos de classificação e fluxo de análise preditiva de dados usando o KDD em dados de vítimas fatais.

A quinta seção é composta pelas considerações finais.

2 REVISÃO DE LITERATURA

A revisão apresentada trata dos temas violência doméstica, gestão da informação, recuperação da informação e descoberta de conhecimento em bases de dados que é subdividida em: etapa de data mining, tarefas de mineração de dados e regras de classificação.

A qualidade de uma pesquisa científica pode ser garantida por muitos fatores, dentre estes, a revisão de literatura. Segundo Galvão e Ricarte (2019, p. 58) em um artigo sobre o tema, propõem a primeira reflexão a ser considerada aos pesquisadores que iniciam suas carreiras acadêmicas: “Revisar a literatura é atividade essencial no desenvolvimento de trabalhos acadêmicos e científicos.”

2.1 VIOLÊNCIA DOMÉSTICA

A violência pode ser definida de diversas maneiras, dentre elas a que mais se adere ao que essa pesquisa pretende explicar é a da publicação do relatório mundial sobre violência e saúde da Organização Mundial de Saúde (OMS):

The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation. (OMS, 2002).

Por sua vez, a Lei Maria da Penha em seu artigo 5º, descreve que “configura violência doméstica e familiar contra a mulher qualquer ação ou omissão baseada no gênero que lhe cause morte, lesão, sofrimento físico, sexual ou psicológico e dano moral ou patrimonial” (BRASIL, 2006). Esta lei, é a mais importante no país para proteção da mulher contra violência e foi concebida, segundo a própria lei para criar “mecanismos para coibir a violência doméstica e familiar contra a mulher [...]” (BRASIL, 2006).

Entre os artigos pesquisados para essa dissertação foram encontrados conceitos de violência doméstica bem parecidos com o da Lei Maria da Penha. Nas bases de dados desse trabalho o atributo que representa a natureza jurídica contém naturezas dessa lei e de outras que se referem a condutas criminosas. Essas bases foram geradas a partir dos sistemas BOU e SCOL, nesses, as naturezas jurídicas se

baseiam no ordenamento legal nacional, como o caso da Lei Maria da Penha e Código Penal Brasileiro. Como exemplo, o BOU tem uma natureza “Lesão corporal – Violência doméstica” e uma natureza “Lesão corporal” que é mais antiga. A Lesão corporal mais antiga está em uma lei bem anterior a Lei Maria da Penha, portanto, foi criada a nova natureza de Lesão corporal ligada à lei que trata da violência doméstica para diferenciá-las.

A Lesão corporal mais antiga no sistema de ocorrências BOU se refere ao artigo 129 do código penal (BRASIL, 1940). Este decreto-lei descreve a conduta criminal no seu caput, porém, acrescenta mais tempo de pena para casos de maior gravidade. Importante destacar a existência dessa gradação quanto a esse crime. A gradação acrescenta detalhes a conduta principal da lei e agrava as penas à medida que a conduta se torna mais destrutiva. As condutas mais destrutivas que agravam as penas são: lesão corporal de natureza grave, lesão corporal de natureza gravíssima e lesão corporal com resultado morte.

Apesar dos termos “lesão de natureza grave” e “lesão de natureza gravíssima” não existirem na lei original, esses termos foram adicionados como naturezas jurídicas no sistema BOU para representar essa diferença de gravidade na conduta e a diferença nas penas aplicáveis.

No Código Penal, o artigo 129 aumenta a pena da lesão corporal quando (no BOU é a lesão corporal de natureza grave) essa resulte à vítima uma ou mais dessas alternativas: “Incapacidade para as ocupações habituais, por mais de trinta dias; perigo de vida; debilidade permanente de membro, sentido ou função; aceleração de parto”. A pena aplicável à lesão corporal sem nenhuma dessas condutas descritas é de reclusão de três meses a um ano, se o fato se enquadrar em qualquer uma das alternativas a pena é modificada para reclusão de um a cinco anos.

A pena é aumentada também quando (no BOU é a lesão de natureza gravíssima) o resultado da ação à vítima é uma ou mais dessas alternativas: “Incapacidade permanente para o trabalho; enfermidade incurável; perda ou inutilização do membro, sentido ou função; deformidade permanente; aborto”. A pena aplicável para qualquer dessas condutas é reclusão de dois a oito anos. A lesão corporal seguida de morte tem a pena aplicável de reclusão de quatro a doze anos.

Da mesma maneira que foi preciso separar as condutas da lesão corporal do código penal brasileiro dependendo de seu potencial destrutivo, a Lei Maria da Penha também cria uma separação, porém, essa separação foi redigida segundo ao tipo de

violência doméstica contra mulher no seu 7º artigo. Os cinco tipos são vistos no QUADRO 1.

QUADRO 1 – TIPOS DE VIOLÊNCIA DOMÉSTICA E FAMILIAR CONTRA A MULHER

Física	Conduta que ofenda a integridade ou saúde corporal.
Psicológica	Conduta que cause dano emocional e diminuição da autoestima, [...] controlar ações, comportamentos, humilhação, manipulação, mediante ameaça, constrangimento, limitar direito de ir e vir, violação de intimidade [...] ou qualquer outro meio que lhe cause prejuízo à saúde psicológica.
Sexual	Conduta que a constranja a presenciar, a manter ou a participar de relação sexual não desejada, mediante intimidação, ameaça, coação ou uso da força; induzir comercialização ou utilizar a sexualidade, impedir o uso método contraceptivo ou forçar ao matrimônio, forçar gravidez, aborto ou prostituição [...] ou que limite ou anule o exercício de seus direitos sexuais.
Patrimonial	Conduta que configure retenção, subtração, destruição parcial ou total de seus objetos, instrumentos de trabalho, documentos pessoais, bens, valores e direitos ou recursos econômicos, incluindo os destinados a satisfazer suas necessidades.
Moral	Conduta que configure calúnia, difamação ou injúria.

FONTE: adaptado de Brasil (2006).

A maioria das publicações da revisão de literatura para essa pesquisa contribuíram com conceitos muito similares ao 7º artigo da Lei Maria da Penha para os tipos de VD, com poucas exceções: no estudo conduzido por Souto et al. (2015), a violência patrimonial apareceu como violência financeira principalmente contra mulheres idosas; por sua vez, Derrick, Testa e Leonard (2014) estudaram a violência verbal sendo equivalentes aos tipos de violência psicológica e/ou moral vistos no QUADRO 1.

Foi observado que na grande maioria das publicações incluídas na pesquisa, os pesquisadores coletaram dados de VD contra mulher focada em um ou dois tipos de violência descritos no artigo 7º, sendo a forma mais comum nessas pesquisas, a violência física.

Salvo pequenas diferenças de nomenclatura dos tipos de VD publicados por diversos órgãos que a estudam e monitoram, é possível atribuir qualquer ação violenta pesquisada nos artigos do corpo de pesquisa dessa dissertação nesses cinco tipos de violência doméstica do artigo 7º da Lei Maria da Penha.

Apesar da Lei Maria da Penha ser uma conquista das mulheres é possível admitir que um homem, criança ou idoso possa sofrer desse mesmo tipo de violência, sendo então, foram acrescentadas as proteções previstas nessa lei a artigos de outros dispositivos legais como o Código penal.

Com um possível efeito da conquista social pela Lei específica sobre VD, em 2021, foi acrescentado o inciso 13 ao artigo 129 do código penal brasileiro, este descreve uma nova conduta ao crime de lesão corporal: “Se a lesão for praticada contra a mulher, por razões da condição do sexo feminino, [...]”. Ao mesmo código acrescentou-se ainda o artigo 147-B, que define em seu *caput*:

Causar dano emocional à mulher que a prejudique e perturbe seu pleno desenvolvimento ou que vise a degradar ou a controlar suas ações, comportamentos, crenças e decisões, mediante ameaça, constrangimento, humilhação, manipulação, isolamento, chantagem, ridicularização, limitação do direito de ir e vir ou qualquer outro meio que cause prejuízo à sua saúde psicológica e autodeterminação. (BRASIL, 2021).

À Lei Maria da Penha nesse mesmo ano, 2021, foi acrescido o artigo 12-C que determinava que:

Verificada a existência de risco atual ou iminente à vida ou à integridade física ou psicológica da mulher em situação de violência doméstica e familiar, ou de seus dependentes, o agressor será imediatamente afastado do lar, domicílio ou local de convivência com a ofendida. (BRASIL, 2021).

As modificações do Código Penal brasileiro e Lei Maria da Penha em 2021 foram dadas pela Lei 14.188 (BRASIL, 2021).

Quanto ao feminicídio, qualificador do crime de homicídio, este foi estabelecido pela lei 13.104, acrescentando ao artigo 121 do Código Penal a nova conduta. Assim se estabeleceu que o homicídio de uma mulher se torna feminicídio quando ocorre: “contra a mulher por razões da condição de sexo feminino: Considera-

se que há razões de condição de sexo feminino quando o crime envolve, violência doméstica e familiar; menosprezo ou discriminação à condição de mulher.” (BRASIL, 2015).

2.2 GESTÃO DA INFORMAÇÃO

A palavra “gestão” tem seu fundamento na ciência da administração. Só se administra algum recurso quando ele é importante, tem valor e pode dar alguma vantagem, do contrário, as ciências administração, economia e contabilidade nem existiriam, pois os recursos seriam ilimitados e tê-los não traria qualquer vantagem.

Quando tratamos dos recursos à disposição, um deles tem se tornado cada dia mais valioso, o Conhecimento. Empresas de todos os ramos de criação de produtos tem seus projetos e protótipos completamente guardados de qualquer pessoa que não seja da alta diretoria ou não faça parte do projeto.

Para chegar ao conhecimento é necessário ter passado por passos anteriores, dados e informação (DAVENPORT, 1998). A definição pode ser vista no QUADRO 2.

QUADRO 2 – DADOS, INFORMAÇÃO E CONHECIMENTO

Dados	Informação	Conhecimento
Simple observações sobre o estado do mundo • Facilmente estruturados • Facilmente obtido por máquinas • Frequentemente quantificado • Facilmente transferível	Dados dotados de relevância e propósito • Requer unidade de análise • Exige consenso em relação ao significado • Exige necessariamente a mediação humana	Informação valiosa da mente humana • De difícil estruturação • De difícil captura em máquinas • Frequentemente tácito • De difícil transferência

FONTE: adaptado de Davenport (1998, p. 18).

O maior motivo de se gerenciar dados, informações e conhecimento é apoiar as tomadas de decisão, como dito por McGee e Prusak (1994, p. 24), “estamos cercados de dados, e as organizações estão repletas de dados que poderiam se tornar informação valiosa para algum usuário diante de um problema decisório”. Certo do valor que dados, informação e conhecimento têm para as instituições, gerenciar é

fundamental. McGee e Prusak (1994, p. 23) consideram que a informação “representa uma classe particular dentre esses outros tipos de ativo. As diferenças decorrem do próprio potencial da informação assim como do desafio de administrá-la ou gerenciá-la.”

A informação nas organizações tem um fluxo e para Beal (2008, p. 29) é descrito como: “identificação de necessidades e requisitos; obtenção; tratamento; armazenamento; distribuição; uso; descarte”.

No caso específico da segurança pública, o descarte não é utilizado devido às peculiaridades dos sistemas que coletam dados.

A forma de recuperar a informação pode melhorar ou piorar o processo de gestão da informação, portanto, o próximo tópico trata desse fator importante.

2.3 RECUPERAÇÃO DA INFORMAÇÃO

Para um dado se tornar uma informação é necessário a mediação humana, criar contexto e relacioná-lo. Na etapa de tratamento da informação Beal (2008) descreve fatores muito importantes na recuperação da informação: adaptação da informação aos requisitos do usuário, classificação temática da informação, classificação da informação quanto a segurança. A autora também ressalta que a classificação da informação não é imutável. Fato que ocorre na categorização de informações que o CAPE faz gestão como modificações em leis, criação de novos campos, criação de novas leis podem causar mudança na estrutura dos dados e a reclassificação de determinada informação. Exemplo de modificação de estrutura de dados foi a inclusão do qualificador de crime de homicídio nominado de “feminicídio”, esse mudou a maneira de interpretar as informações e levou a reestruturação dos processos de análise dos homicídios, por sua vez, modificou o método de classificação da informação dessa natureza jurídica.

A recuperação da informação também tem seus resultados associados à capacidade de retorno dos mecanismos de busca e localização da informação que a organização utiliza. Esta é relacionada a parte do fluxo da informação nominada distribuição segundo Beal (2008) em seu tópico “Aspectos relacionado à distribuição da informação”.

Uma recuperação da informação bem planejada tem maiores chances de possibilitar a aplicação de métodos como o KDD para apoiar decisões, apresentados

no próximo tópico. Nas etapas do KDD, principalmente as duas primeiras dessas, a forma como as informações são recuperadas impactará diretamente na aplicação de métodos de Mineração de Dados, podendo causar a impossibilidade de uso de alguns desses. A impossibilidade de uso de alguns métodos pode alterar a capacidade de respostas às questões necessárias para a tomada de decisão, portanto, o planejamento de recuperação deve levar em conta o que se quer responder e os métodos que serão usados para buscar a resposta.

2.4 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (KDD)

A descoberta de conhecimento em base de dados se originou da junção de várias outras áreas de conhecimento. Goldshmidt e Passos (2005) destacam algumas dessas áreas: estatística, descobrimento de padrões, banco de dados, inteligência computacional e aprendizado de máquina.

O termo “Knowledge Discovery in Databases” (KDD) foi proposto por Piatetsky-Shapiro para um *workshop*. O KDD “tem como objetivo encontrar padrões intrínsecos aos dados nela contidos, apresentando-os de forma a facilitar sua assimilação como conhecimento” (DA SILVA; PERES; BOSCARIOLI, 2016, p.11).

O processo de KDD segundo Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 84) é composto por cinco etapas: seleção, pré-processamento, transformação, *data mining* e interpretação/avaliação. Este mesmo processo descrito pelos autores pode ser visto com pequenas variações de nomenclatura (traduzida) e pode ter o processo de transformação suprimido. Nessa pesquisa o processo de Fayyad é utilizado com as suas cinco etapas.

A etapa Seleção trata da extração dos dados nos bancos de dados, feita através comandos escritos em *Structured Query Language* (SQL).

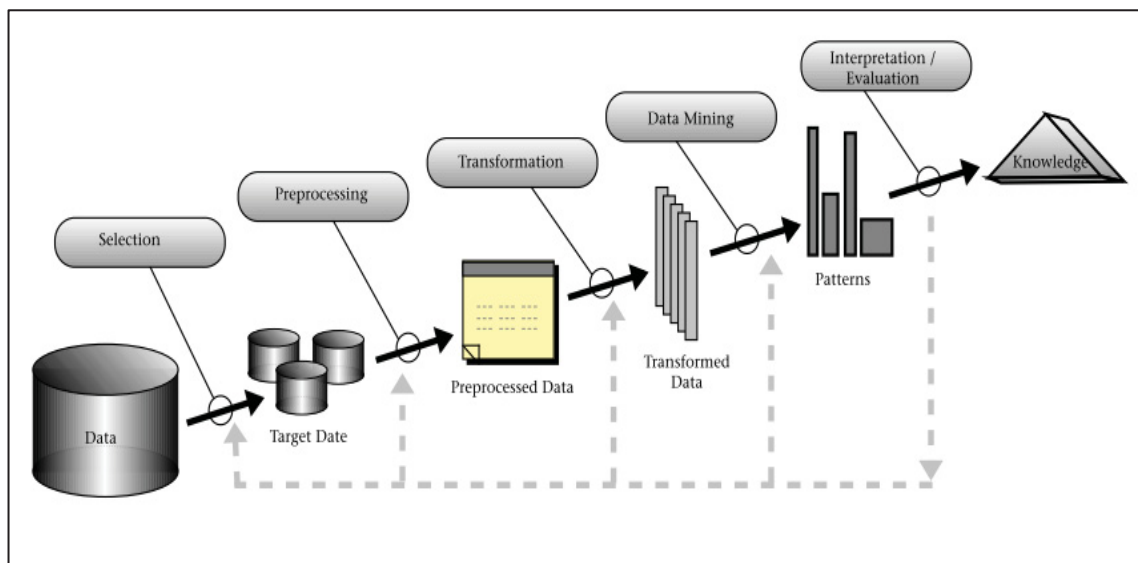
Pré-processamento trata de processamentos dos dados, como cálculos de campos necessários, como determinar a idade de uma pessoa feita através da data de nascimento ou somar alguns atributos agrupando-os por linha que também pode ser feito através de comando SQL.

Transformação é a etapa onde alguns atributos são transformados para a etapa de mineração de dados, como por exemplo, a discretização¹ de atributos contínuos, eliminação de atributos que trazem dados duplicados ou que não são relevantes na etapa de mineração de dados.

Data Mining é a etapa onde as tarefas de mineração são aplicadas, já foi considerado erroneamente como sinônimo para KDD, devido a sua importância será tratado no tópico 2.4.1.

Interpretação/Avaliação é uma etapa fundamental pois trata da avaliação e interpretação dos resultados encontrados. Se avalia no resultado encontrado a utilidade dos padrões, o que pode ser levado em conta ao descrever um conjunto de dados, o que pode ser considera anomalia, entre outros exemplos. A interpretação dos resultados nessa etapa depende do conhecimento de estatística e das técnicas utilizadas no processo de *data mining*. Estas etapas podem ser vistas na FIGURA 1.

FIGURA 1 – ETAPAS DO KDD



FONTE: Fayyad; Piatetsky-Shapiro e Smyth (1996).

Além do modelo de KDD proposto por Fayyad, dois modelos que foram criados bem menos ligados à academia, SEMMA e CRISP-DM, se tornaram muito populares entre grandes empresas de diversas áreas. Segundo Azevedo e Santos (2008) o SEMMA foi criado pela *Statistical Analysis System* (SAS) e o *Cross Industry*

¹ Dependendo dos algoritmos de mineração de dados e do objetivo do KDD é necessário transformar valores de atributos contínuos em intervalos de valores.

Standard Process for Data Mining (CRISP-DM) foi criado inicialmente por um consórcio composto por *Daimler Chrysler*, *Statistical Package for the Social Science* (SPSS) e *NCR Systems Engineering Copenhagen*. O CRISP-DM e SEMMA tem similaridades de conceitos, porém o primeiro modelo tem uma etapa a mais. Uma comparação entre os modelos pode ser vista na FIGURA 2.

FIGURA 2 – COMPARAÇÃO ENTRE KDD, SEMMA E CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	-----	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	-----	Deployment

FONTE: Azevedo e Santos (2008).

As etapas parecem muito similares, entretanto, SEMMA E CRISP-DM adaptaram alguns conceitos às necessidades de seus criadores. A etapa Pré KDD e Business understanding (CRISP-DM) são correspondentes, porém, da criação do modelo CRISP-DM antes de avançar etapas é necessário descrever o contexto e propósito da aplicação do modelo, o SEMMA não tem etapa similar para esta.

Etapas Selection (KDD) e Sample (SEMMA) são bem similares, a Data Understanding (CRISP-DM) compreende as duas primeiras citadas juntamente com as etapas *Pre processing* (KDD) e *Explore* (SEMMA).

Transformation, *Modify* e *Data preparation*; *Data Mining*, *Model*, *Modeling* e; *Interpretation*, *Assessement*, *Evaluation*, tem diferença de nomenclatura, mas são etapas praticamente idênticas.

2.4.1 Data Mining

A etapa do KDD, *data mining*, segundo Thuraisingham² (1999 apud Côrtes et al., 2002, p. 1), “é um processo de proposição de várias consultas e extração de

² THURASINGHAM, B. M. **DATA MINING**: Technologies, Techniques, Tools, and Trends. Boca Raton: CRC Press LLC, 1999.

informações úteis, padrões e tendências, frequentemente desconhecidos, a partir de grande quantidade de dados armazenadas em bancos de dados”.

Witten, Frank e Hall (2011) tratam do conceito de forma parecida, porém adicionam informações sobre a descoberta, sendo essa a que proporciona vantagem e automatização dos procedimentos:

Data mining is defined as the process of discovering patterns in data [...] automatic or semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic one. The data is invariably present in substantial quantities. (WITTEN; FRANK; HALL, 2011).

A descoberta de padrões cria a vantagem de se ter conhecimento sobre um objeto de análise. Essa vantagem se estabelece quando o que se descobre pode modificar as tomadas de decisões.

2.4.2 Tarefas de mineração de dados

Quanto as tarefas de mineração de dados, Da Silva, Peres e Boscarioli (2016, p. 12) as classificam em preditivas e descritivas. As tarefas descritivas têm como métodos: agrupamento, sumarização, modelagem de dependências e detecção de desvios; as tarefas preditivas têm como métodos: classificação e regressão.

Tarefas preditivas usam os valores dos atributos descritivos para prever valores futuros ou desconhecidos de outros atributos de interesse. Já as tarefas descritivas têm o objetivo de encontrar padrões que descrevem os dados de maneira que o ser humano possa interpretar. (DA SILVA; PERES; BOSCARIOLI, 2016, p. 12).

O fluxo de análise de dados preditiva de VD proposto, utiliza as regras de classificação para encontrar padrões ainda desconhecidos. Os algoritmos PRISM e CN2 foram selecionados por terem capacidade de induzir regras com atributos nominais e terem como base o algoritmo *ID3* (QUINLAN, 1987), originalmente criado para árvores de decisões. Esses algoritmos serão explicados em detalhes no item 2.7.1 e 2.7.2.

2.5 REAMOSTRAGEM (RESAMPLING)

O desbalanceamento de classe em bases de dados é uma característica de algumas dessas bases. Pesquisas com base de dados da área de saúde de Lee (2014) e de transações financeiras de Lima (2016) relatam tal característica como um complicador para a mineração de dados. Como um exemplo, uma base de dados de transações financeiras com 100 observações é provável que menos de 3% sejam fraude. Ao submeter essa base a um algoritmo de árvore de decisões, muito provavelmente o resultado seja apenas um nó sugerindo que não existe fraude nessas transações pois a classe majoritária é composta de observações financeiras feitas de forma legítima. O caso específico de vítimas de feminicídio com histórico de VD desse estudo impõem desafios parecidos com os de Lee e Lima na etapa de mineração de dados.

A busca por soluções para problemas de desbalanceamento de classe foi uma preocupação de pesquisadores da área acadêmica, esses produziram soluções baseando-se na estatística. Lee (2014) em sua pesquisa na área de KDD com dados de saúde chegou à conclusão de que, a aplicação de técnicas de *resampling* melhoraram as medidas de avaliação (acurácia, área sob a curva etc.), predição de valores positivos e aumentam o poder de explicação em algoritmos como árvores de decisão, *random forrests* e *generalized boosted trees*.

Então, é indicado o uso de alguma estratégia de *resampling* ao realizar o KDD em uma base de dados com classes muito desbalanceadas, principalmente quando na mineração de dados, os dados são aplicados a um algoritmo de classificação (LIMA, 2016).

2.5.1 Oversampling

Uma estratégia de *resampling* que pode ser empregada em bases com classes desbalanceadas é o *oversampling*. Nessa ocorre a replicação artificial de instâncias para a classe minoritária mantendo as instâncias originais dessa classe (LIMA, 2016). Em suma, o equilíbrio é criado aleatoriamente nas instâncias da classe minoritária até que o número de instâncias dessa seja igual ou bem próximo do número de instâncias da classe majoritária. Um exemplo seria uma base de dados como a usada por Lee (2014) para determinar a diabetes em pacientes, somente 3%

apresentavam o valor positivo para tal. Utilizando *oversampling* as linhas com diagnóstico positivo para diabetes seriam replicadas aleatoriamente até que sua quantidade fosse igual as linhas com diagnóstico negativo.

A crítica feita a essa estratégia é que os resultados tendem a um super ajuste (*overfitting*), prejudicando os resultados. O super ajuste geralmente melhora as medidas objetivas de avaliação de modelos preditivos, mas podem contaminar os resultados tornando-os mais complexos e menos generalizados (CHAWKA, 2005; LIMA, 2016).

2.5.2 Undersampling

A estratégia de *undersampling* tem a lógica inversa à anterior, *oversampling*. São mantidas as instâncias da classe minoritária e são retiradas aleatoriamente instâncias da classe majoritária para compor uma nova base amostral. O risco do uso dessa estratégia é que o algoritmo usado na tarefa não selecione instâncias com dados fundamentais da classe majoritária para o aprendizado do algoritmo de mineração de dados (LIMA, 2016).

A questão de aleatoriedade da escolha em muitos casos pode ser um complicador, entretanto, a estratégia *undersampling* trabalha com dados reais, não fictícios como caso da estratégia anterior. Por melhor que sejam as medidas de avaliação objetivas em um modelo preditivo com uso de *oversampling*, o resultado pode ser fruto do efeito da replicação da classe minoritária.

2.6 SELEÇÃO DE ATRIBUTOS

Na etapa de mineração é comum que as bases de dados tenham uma quantidade numerosa de atributos que se não forem removidos tendem a dificultar o processo de aprendizagem dos algoritmos. A busca por soluções para esse problema de dimensionalidade foi iniciada na estatística. No entanto, essa solução não é trivial e definitivamente não é única (BORGES, 2006). Muito foi produzido nessa área desde o fim dos anos noventa usando abordagens diferentes.

A lógica da seleção de atributos está apoiada em encontrar um conjunto de atributos que melhore o desempenho dos algoritmos sem perder qualidade nos resultados.

Segundo Lee (2005), as abordagens de seleção de atributos podem ser subdivididas em filtro, wrapper, embedded e híbrida. Com exceção da híbrida, essas abordagens são descritas no QUADRO 3.

QUADRO 3 – ABORDAGENS DE SELEÇÃO DE ATRIBUTOS

Abordagem	Descrição	Autor
Filtro	Essa abordagem não usa algoritmo de aprendizado para avaliar subconjuntos de dados. Ela avalia os conjuntos de dados através de medidas independentes como: de informação, de distância, de dependência, de consistência e de precisão. Exemplos dessas medidas são: entropia, correlação, informação mútua, chi-quadrado.	(LIMA, 2016; MACEDO, 2012; SOUZA, 2017)
Wrapper	Essa abordagem utiliza um algoritmo de mineração de dados e usa o desempenho como critério de avaliação. Portanto, é necessário estabelecer previamente o algoritmo de mineração de dados. Essa abordagem busca sempre melhorar o desempenho dos algoritmos de mineração, pois ele busca quase que incessantemente o melhor conjunto de atributos, o que o leva a ter um custo computacional bem maior que a abordagem filtro.	(SOUZA, 2017)
Embedded	Também conhecida como embutida, nessa abordagem a seleção de atributos é realizada no algoritmo de mineração de dados que será usado na descoberta de padrões. Portanto, essa é realizada no processo de indução do classificador deixando de ser parte da etapa de pré-processamento.	(LIMA, 2016)

FONTE: O Autor (2022).

Segundo Liu e Yu (2005), a abordagem é considerada híbrida quando se usa em conjunto a abordagem filtro e wrapper.

Na seleção de atributos das bases de dados aqui analisadas, considerando a replicabilidade, softwares utilizados e testes realizados por Souza (2017), se empregou a abordagem filtro com quatro técnicas: CFS, Info Gain, Gain Ratio e ReliefF.

2.6.1 Correlation-based feature selection (CFS)

Esse método, utilizando o conceito de relevância e de redundância, avalia a capacidade de um subconjunto de dados representar todo um conjunto de dados. A relevância é medida pelo cálculo de correlação de um atributo em relação a uma classe, já a redundância é medida calculando a correlação desse mesmo atributo em relação aos demais atributos de um subconjunto. De fato, o propósito do CFS é escolher o que tem mais poder de predição diminuindo a dimensionalidade, portando, elimina a redundância (LIMA, 2016).

Para um atributo ser escolhido para o subconjunto pelo CFS, esse deve ter a correlação com a classe maior que a correlação mais alta entre ela e qualquer uma das feições já selecionadas (HALL, 2000).

Essa técnica analisa os atributos individualmente em relação à classe e em relação ao subconjunto. Isso o torna diferente da maioria dos métodos de seleção de atributos pois esses avaliam os atributos se baseando em um único conceito, nesse se considera a relevância e a redundância.

2.6.2 Info Gain

O *Info Gain*, baseado em filtragem, opera com a avaliação univariada de um atributo estabelecendo um ranking. Esse, calcula o mérito de um atributo baseado no ganho de informação que ele proporciona no conjunto. Para determinar o ganho de informação o algoritmo age semelhante a métrica de informação mútua baseando-se no conceito de entropia, o que de certa forma o aproxima da forma de calcular do CFS. A redução da entropia é o que determina o ganho de informação de um atributo a um conjunto (LIMA, 2016).

A interpretação dos resultados de uma técnica como essa é bem intuitiva, o que a torna uma forte candidata em uso para selecionar atributos. No entanto, muitas críticas foram feitas a essa técnica por ser insensível a alguns fatores importantes, como o caso de calcular o maior ganho para um atributo de identificação sequencial. Mesmo que o cálculo esteja correto, o atributo de identificação deve atrapalhar o processo de descoberta dos algoritmos utilizados na mineração de dados.

2.6.3 Gain Ratio

O Gain Ratio segundo Lima (2016), é a versão ponderada da técnica anteriormente descrita, Info Gain. Criado para resolver limitações do Info Gain seleciona atributos com maiores valores distintos. Então, a técnica propõe-se a buscar por “atributos que maximizam o ganho da informação, enquanto minimizam o número de valores de um atributo.” (LIMA, 2016, p. 21).

A ideia principal dessa técnica é suavizar a diferença entre valores do atributo e o seu ganho, o que o torna mais sensível a identificar atributos com a característica do identificador sequencial do que o algoritmo anterior.

O uso de técnicas com ganho de informação como essa e a anteriormente citada são mais apropriadas para bases de dados com poucos atributos.

2.6.4 ReliefF

Segundo Lee (2005), o algoritmo relief trabalha através de amostras aleatórias de exemplos de um conjunto de dados da localização do vizinho (distância) mais próximo da mesma classe e da classe oposta, trabalhando somente com duas classes. O ReliefF, versão estendida do relief, trabalha com múltiplas classes e consegue lidar com ruídos de dados (LEE, 2005).

O ReliefF e sua versão anterior, calcula a relevância dos atributos como o modelo de aprendizagem baseada em instância. A relevância (estatística) no reliefF é calculada através da distância de *manhattan* (LEE, 2005; NEVES, 2003).

2.7 REGRAS DE CLASSIFICAÇÃO

As regras de classificação buscam correspondência entre atributos que normalmente não são verificadas de forma intuitiva lendo uma pequena base de dados. Essas geralmente são utilizadas em bases de dados de sistemas especialistas com intuito de extrair conhecimento novo pois os especialistas podem interpretá-las com facilidade (HALMENSCHLAGER, 2002).

Algoritmos com essa finalidade tem em sua implementação algoritmos de árvores de decisões adaptados. O conceito internamente é o mesmo das árvores, mas

a forma com que se o algoritmo entrega o resultado é diferente. As árvores de decisão apresentam um resultado gráfico com as conexões do conhecimento criado, no entanto, as regras são apresentadas de forma independente, ou seja, são escritas na maioria dos algoritmos em linguagem mais próxima à natural (geralmente em inglês).

Uma regra gerada em um desses algoritmos para uma base de dados meteorológicos poderia ser representada dessa forma: *IF* céu nublado *AND* umidade do ar = 97% *THEN* chove. Nesse caso o *IF* representa a condição (céu nublado), *AND* adiciona uma segunda condição conjunta (umidade do ar igual a 97%) e *THEN*, então (chove).

2.7.1 Algoritmo PRISM

O PRISM foi criado por J. Cendrowska em 1987, Vasconcelos (2002) ao tratar do método PRISM afirma que este é “baseado no algoritmo de indução de regras ID3. A motivação na sua criação foi de superar as limitações na representação de árvores de decisão. PRISM produz regras modulares (independentes) que são de mais fácil compreensão.” (VASCONCELOS, 2002, p. 33).

Em MD com bases de dados reais, a capacidade de representação das árvores de decisões pode dificultar a compreensão quando o número de nós e galhos é muito extenso, então, o PRISM foi desenvolvido como alternativa às árvores de decisão.

As regras geradas por esse algoritmo são no formato “*IF* regra *THEN* classe predita” facilitando a leitura e compreensão da descoberta.

Diferentemente do método dividir-e-conquistar, nos quais se baseiam os algoritmos de árvore de decisão, a abordagem adotada pelo PRISM é ‘covering’ e ‘bottom-up’. O PRISM consiste em: dada uma massa de dados como entrada ao algoritmo este gerará somente regras ‘corretas’ (a regra gerada cobrirá apenas as instâncias que satisfazem as condições que aparecem no antecedente da regra). (VASCONCELOS, 2002, p. 34).

A cobertura condicional é baseada em probabilidade estatística e a abordagem de baixo para cima. Após o algoritmo gerar uma regra, esse remove todas as instâncias cobertas pela regra recém-criada do conjunto de treinamento e reinicia o processo de geração de regras. O algoritmo induz regras em bases com atributos categóricos, se apoia na entropia como medida interna e modificou conceitos existentes no ID3 para melhorar a performance de classificação buscando somente

valores relevantes de atributos e não o atributo mais relevante em geral (CENDROWSKA, 1987).

2.7.2 Algoritmo CN2

O CN2 foi criado por Clark e Niblett (1989) adaptando ideias de dois algoritmos, ID3 de Quinlan (1987) e AQ de Michalski³.

O algoritmo ID3 adaptado (*assistant*) possibilita minerar dados ruidosos devido a abordagem de crescimento de cima para baixo. Esse usa a entropia para selecionar o melhor nó e determinar sua posição na árvore. Também, o super ajuste (*overfitting*) é evitado nesse algoritmo interrompendo o crescimento da árvore quando nenhuma informação gerada é significativa (CLARK; NIBLETT, 1989).

O funcionamento do CN2 consiste em

[...] dois procedimentos: um mecanismo de pesquisa por boas regras e um mecanismo de controle para executar repetidamente esta pesquisa. O CN2 usa uma função de avaliação para selecionar o atributo que ajuda a distinguir uma classe [...] das outras, e adiciona o teste resultante em uma regra conjuntiva. Este processo é repetido até que a regra exclua todos os exemplos de outras classes e remova os exemplos da classe que a regra cobre, repetindo este processo nos exemplos de treinamento remanescentes. (HALMENSCHLAGER, 2002, p. 56).

Uma vantagem do CN2 para o PRISM é a possibilidade de minerar em dados contínuos além dos categóricos. Tal vantagem pode ampliar as possibilidades de analisar dados.

2.8 AVALIAÇÃO DE MODELOS PREDITIVOS DE CLASSIFICADORES

A última etapa do modelo de KDD proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996) e a junção dos termos “Interpretação” e de “Avaliação”. O processo de avaliação de um modelo preditivo se orienta em métricas objetivas. Segundo Da Silva, Peres e Boscarioli (2016), modelos baseados em tarefas de classificação utilizam como sua medida mais comum a acurácia, ou seja, a taxa de classificação correta.

³ MICHALSKI, R. S. On the quasi-minimal solution of the general covering problem. **Proceedings of the Fifth International Symposium on Information Processing**, p. 125-128, 1969. Bled, Yugoslavia.

Os algoritmos de MD de tarefas de classificação além da acurácia também apresentam outras medidas importantes no processo de avaliação. Por mais que a acurácia seja uma medida de simples entendimento e avalia um fator importante do modelo, essa deve ser avaliada em conjunto com outras medidas.

A observação da matriz de confusão, é muito importante na avaliação modelos preditivos baseados em classificação. Essa é a base para o cálculo da maioria das medidas consideradas na avaliação. A observação dessa matriz pode evidenciar a alta capacidade de prever ou o contrário, a incapacidade do modelo em prever a classe para os dados submetidos a um algoritmo. Um exemplo de matriz de confusão de uma classe binária pode ser visto no QUADRO 4.

QUADRO 4 - MATRIZ DE CONFUSÃO DE CLASSE BINÁRIA

		Classe Predita ($f(x)$)	
		positivo	negativo
Classe Esperada (y)	positivo	Verdadeiros positivos (VP)	Falsos negativos (FN)
	negativo	Falsos positivos (FP)	Verdadeiros negativos (VN)

FONTE: adaptado de Da Silva; Peres; Boscarioli (2016).

Os Verdadeiros positivos representam a classificação correta na classe positiva e os Falsos positivos representam a classificação incorreta na classe positiva. Os Verdadeiros negativos representam a classificação correta na classe negativa e os Falsos negativos representam a classificação incorreta na classe negativa (DA SILVA; PERES; BOSCARIOLI, 2016).

Dentre muitas medidas para avaliação de modelos preditivos, em especial as com classificadores, algumas são comumente empregadas. A acurácia, sensibilidade, especificidade, precisão e F1-Score. O QUADRO 5 apresenta uma breve descrição e a fórmula de cada medida.

QUADRO 5 - MEDIDAS PARA AVALIAÇÃO DE MODELO PREDITIVO COM CLASSIFICADORES

Medida	Acurácia	F-Score	Sensibilidade	Especificidade	Precisão
Descrição	(Accuracy ou ACC) é considerada uma das métricas mais simples e importantes. Ela avalia simplesmente o percentual de acertos, ou seja, ela pode ser obtida pela razão entre a quantidade de acertos e o total de entradas.	F-measure, F-score ou Score F1 é uma média harmônica calculada com base na precisão e na sensibilidade.	(Também conhecida como Recall ou Revocação). Essa métrica avalia a capacidade do método de detectar com sucesso resultados classificados como positivos.	Avalia a capacidade do método de detectar resultados negativos.	(Precision) é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos.
Fórmula	$(VP+VN)/(VP+FN+VN+FP)$	$2*((\text{precisão}*\text{sensibilidade})/(\text{precisão}+\text{sensibilidade}))$	$VP/(VP+FN)$	$VN/(VN+FP)$	$VP/(VP+FP)$

FONTE: adaptado de Mariano (2021).

Essas medidas possibilitam avaliar com objetividade, ajudam a identificar a capacidade de acertar, detectar o sucesso na classificação positiva, detectar resultados negativos, determinar a precisão pelos positivos e verificar a média harmônica entre precisão e sensibilidade.

Como exemplo, se uma classificação atinge um valor de acurácia 0,775, um valor relativamente próximo do classificador perfeito (1), uma sensibilidade de 0,850, uma especificidade de 0,7, precisão de 0,739 e F-Score de 0,790, significa que mesmo que a acurácia seja um bom valor, o modelo tem maior facilidade ao detectar valores positivos e maior dificuldade para detectar os valores negativos. O classificador erra muito mais em uma classe, é necessário estudar as causas dessa assimetria.

Portanto, a matriz de confusão e as medidas de avaliação são fundamentais para se avaliar um classificador, mais ainda quando se compara dois ou mais classificadores.

2.8.1 Validação cruzada

Os algoritmos de tarefas de classificação podem utilizar uma avaliação interna conhecida como validação cruzada. Comumente se encontra na literatura especializada a validação cruzada com 10 partes ou 10-folds.

O procedimento da validação cruzada realizada em uma base consiste em dividir o conjunto de treinamento em um número de subconjuntos (NEVES, 2003, p. 117). Se o valor determinado for 10, o algoritmo realiza o treinamento com 9 partes e o resultado é avaliado com o subconjunto não utilizado no treinamento.

Halmenschlager (2002) tratando da validação cruzada explica que,

Este método possui uma tendência pessimista, dado que cada um dos classificadores utilizados para estimativa é construído apenas com um subconjunto do treinamento. Já o classificador final, que será aquele que obtiver a melhor taxa individual [...] (HALMENSCHLAGER, 2002, p. 65).

A tendência é de pessimismo pois, na prática, o uso da validação costuma diminuir os valores das medidas de avaliação como a acurácia, porém, aumenta a confiabilidade da descoberta pois se baseia no subconjunto mais qualificado.

2.9 ESTATÍSTICA DESCRITIVA

A estatística segundo Crespo (1986, p. 13) “é uma parte da Matemática Aplicada que fornece métodos para a coleta, a organização, a descrição, a análise e a interpretação de dados para a tomada de decisões.”

Para Freund (2000), a estatística descritiva “compreende o manejo dos dados para resumi-los ou descrevê-los, sem ir além, isto é, sem procurar inferir qualquer coisa que ultrapasse os próprios dados”. (FREUND; SIMON, 2000 apud DIEHL et al. 2007).

A estatística descritiva tem a lógica de generalizar para explicar um conjunto de dados. Medidas como média, moda, desvio padrão, variância, dispersão e coeficiente de variação são empregadas para descrever o conjunto de dados das bases SCOL, BOU e Naturezas quando o atributo permitir tal emprego. Nos demais casos se emprega a análise de frequência dos atributos.

Um fato importante de se ressaltar é que a estatística descritiva sempre está presente em estudos científicos com abordagem quantitativa, ora como único método ou em conjunto com outros métodos numéricos.

2.10 VIOLÊNCIA DOMÉSTICA E ANÁLISE DE DADOS

Na busca por produção científica sobre os dois temas relacionados, violência doméstica e análise de dados, foram selecionadas duas bases de publicações científicas multidisciplinares, Scopus e Web of Science (WoS).

Optou-se pelo uso da revisão integrativa de literatura para orientar a pesquisa em bases de publicação científica sobre os temas dessa dissertação. A revisão integrativa de literatura “é um método que tem como finalidade sintetizar resultados obtidos em pesquisas sobre um tema ou questão de maneira sistemática, ordenada e abrangente.” (ERCOLE; MELO; ALCOFORADO, 2014, p. 9).

A pesquisa resultou em um artigo científico intitulado “Análise de Dados de Violência Doméstica: Revisão Integrativa” que foi apresentado ao Congresso Nacional de Administração (CONAD) 2021 e publicado nos anais⁴ desse congresso.

Utilizou-se a estratégia de busca em português e inglês. A estratégia foi formulada depois de buscas iniciais no Portal CAPES. A busca foi feita nas bases já citadas nos dias 19, 27 e 29 de abril de 2021.

QUADRO 6 – TERMOS DE BUSCA EM PORTUGUÊS E INGLÊS APLICADOS NAS BASES SCOPUS E WEB OF SCIENCE

Título ou Resumo ou Palavras-chave em Português
((“Análise de Dados” OR “Mineração de Dados” OR “Descoberta de Conhecimento em Bases de Dados”) AND “Violência Doméstica”)
Título ou Resumo ou Palavras-chave em Inglês
((“Data Analysis” OR “Knowledge Discovery in Databases” OR “Data Mining”) AND “Domestic Violence”)

FONTE: Da Silva Costa, Tsunoda e Pecini (2021).

A busca com os termos em português não retornou nenhum artigo nas ocasiões. As duas bases retornaram artigos quando foi pesquisado com os termos e inglês. Para aceitação nessa pesquisa foram incluídos somente artigos com acesso livre, dos últimos dez anos, nos idiomas português, inglês e espanhol.

⁴ https://drive.google.com/file/d/1R-y8s_mEP6LrelU9NW2OQq2jnoSWTeLL/view.

Dos 90 artigos retornados pela base WoS e 378 artigos retornados pela base Scopus, 122 atenderam aos critérios estipulados para pesquisa, sendo, 28 pela WoS e 94 pela Scopus.

Foi criado um banco de dados com os dados apontados na pesquisa. O arquivo foi importado do MS Excel⁵ para o software XAMPP e está disponível no site github⁶.

A primeira verificação quanto a existência de artigos científicos em português e inglês, já citada na justificativa, foi confirmada na busca nas bases multidisciplinares. Muitos dos pesquisadores que publicaram seus artigos nas referidas bases têm o português como língua mãe, porém, publicaram em inglês. Artigos escritos em inglês aumentam as chances de alcançar mais membros da comunidade científica internacional, já que esse idioma é usado como segunda língua por pesquisadores.

A pesquisa de análise de dados em VD a partir de 2017 teve um crescimento nas publicações nessas bases, porém ainda não é um tema muito pesquisado. A dificuldade de se conseguir dados de VD pode ser um fator que desincentiva a pesquisa na área, em outra via os dados de violência devem ser protegidos sob as visões legal e moral.

Após a exclusões de artigos sem relação com a análise de dados em VD, duplicados (19 duplicados) e os que não estavam de acordo no quesito qualidade do modelo proposto por Kitchenham e Charters (2007), o corpo de estudo incluído nessa pesquisa totalizou 65 artigos. Da Silva Costa, Tsunoda e Pecini (2021) detalharam tais procedimentos e métodos de inclusão na revisão.

Uma característica da violência doméstica é a revitimização. A revitimização ocorre quando uma vítima sofre diversas vezes a experiência da violência doméstica (MANZANARES; TORRADO TARRÍO; SALGADO, 2011). Aghakhani et al. (2015) verificaram 83% de reincidência de VD contra mulher no Iran, Karystianis et al. (2019) verificaram 45% com mais de um evento e 27% com três ou mais eventos na Inglaterra. Em pesquisa feita por Silwal e Thapa (2020) no Nepal sobre VD contra mulher infértil, constataram a vitimização de 52,63% no primeiro caso de infertilidade e 61,11% no segundo caso. Na base de dados do BOU foi observado que a

⁵ Software da empresa Microsoft com diferentes funcionalidades: planilha de cálculo eletrônica, estatística, planilhas programadas, gerador de dashboard, gerador de gráficos.

⁶ <https://github.com/ChristianSCosta/Violeciadomestica2021ri/blob/main/dbvdresearch.sql> .

revitimização é um fator comum entre as vítimas. Ou seja, as vítimas de VD do estado do Paraná sofrem violência em mais de uma ocasião, similar aos resultados relatados em alguns artigos da revisão integrativa.

2.10.1 Métodos numéricos utilizados em pesquisas de VD

Dos 65 artigos estudados notou-se uso de vários métodos diferentes, tendo destaque para regressão, usada em 27 deles. A Regressão de Cox foi usada por Caprara et al. (2020) e a Regressão de Poisson foi usada por Enriquez-Canto et al. (2020). A regressão de Poisson e de Cox utilizam variável tempo na equação. A regressão linear foi utilizada por Khayat et al. (2017) em um estudo de associação entre violência física e sexual na saúde mental de mulheres.

Os outros métodos de Regressão relatados, 24 usaram Regressão Logística. A Regressão Logística foi relatada em estudos de Fatores de Risco, nesses casos foi descrita como método preditor e foi usada também em estudos de grupos e associação de variáveis externas com a ocorrência de VD (CLARKE et al., 2020; DASGUPTA et al., 2013; DUTTA et al., 2018; FULLER-THOMSON; RAMZAN; BAIRD, 2016; GEBREZGI et al., 2017; HAN; CHOI, 2021; LAEHEEM; BOONPRAKARN, 2015; TANRIVERDI; ÇAPIK; YALÇIN GÜRSOY, 2018).

Verificou-se também o relato do uso da estatística descritiva, 13 no total, sendo que 3 relataram ela como único método, 4 juntamente com regressão, 3 com análise multivariada, 2 com ANOVA e 1 com MANOVA. Apesar de não ter sido encontrado mais relatos do uso de estatística descritiva no corpo de pesquisa, os pesquisadores incluídos na revisão utilizaram a estatística descritiva para descrever a população estudada.

2.11 SOFTWARES

Os softwares têm papel importante no trabalho de pesquisadores, principalmente nos casos de estudos como este que usa mineração de dados. É imprescindível que se escolham softwares que auxiliem na execução dos objetivos propostos no planejamento da pesquisa.

Então, a busca por ferramentas de KDD aqui empregadas levou em consideração alguns fatores importantes como: a capacidade de pré-processar e

transformar dados; os algoritmos pertencentes ao corpo inicial de instalação; a possibilidade de instalar pacotes; a existência de pacotes específicos para cumprir etapas da pesquisa; facilidade ou complexidade no uso; facilidade de exportar resultados.

2.11.1 Software Orange

O Orange é uma ferramenta baseada em Python para mineração de dados desenvolvida no Laboratório de Bioinformática da Faculdade de Ciência da Computação e da Informação, Universidade de Liubliana, Eslovênia. Ele pode ser utilizado por scripts Python como um plug-in ou por meio do *Orange Canvas*, sua interface visual (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014).

Esta ferramenta tem em sua versão 3.31 vários algoritmos implementados em Python para etapas do KDD. A possibilidade de uso do software em formato de fluxograma privilegia a visualização das etapas.

Algoritmos com amplo uso na academia como *Naive Bayes*, *KNN*, *Decision Trees*, *Random Forest*, *SVM*, *Regressão linear*, *Regressão Logística* e *Redes Neurais* estão implementados no conjunto inicial de instalação. Ainda, como algoritmo de indução de regras consta o CN2, mais bem explicado nos itens relacionados a regras de classificação.

O Orange é uma ferramenta livre e permite a instalação de novos pacotes como: séries temporais, educacional (com regressão polinomial e classificação polinomial) e Redes Neurais (outras configurações e redes além do Rede Neural da versão inicial). A possibilidade de instalação de pacotes como estes citados aumenta as possibilidades de emprego de um software como esse em KDD.

Também, esse software tem algumas ferramentas de visualização de dados e exportação de resultados/gráficos em formato *html*, o que possibilita uma visualização bem estruturada dos resultados (utiliza estrutura de tabelas) e armazenar esses em arquivo a cada modificação de parâmetros e/ou execução de algoritmo a pedido do usuário.

2.11.2 Software RStudio

O RStudio é o programa que oferece uma interface que permite manipular o sistema R através de scripts. O sistema R segundo De Andrade (2018),

consiste em um software avançado de estatística compatível com sistemas operacionais como Microsoft Windows e Mac OS X, que, a partir de comandos básicos e funções derivadas, seleciona bases de dados disponíveis em sistemas cadastrados (inclusive internet, bastando, para isso, carregar os dados no próprio sistema), e pode construir uma robusta análise estatística desses dados, sem que seja necessário comprar módulos adicionais para executar funções extras. (DE ANDRADE, 2018, p. 684).

Além da capacidade de se produzir análises estatísticas nesse software também é possível utilizá-lo em etapas de KDD. É possível executar reamostragem, discretização, transformação de dados, realocação de dados, utilizar diversos algoritmos de mineração de dados muito utilizados profissionalmente e academicamente, isso, entre muitas outras aplicações possíveis do R para o KDD.

A diferença em utilizar o R através do RStudio está principalmente nesse estar organizado em um painel composto por padrão em: ambiente (todas as variáveis carregadas), valores carregados na memória, arquivos do diretório principal, o script da programação e a tela do console (da execução dos scripts). A visualização em conjunto possibilita observar muitas informações importantes ao mesmo tempo. Como exemplo, escrever um script e executá-lo em etapas enquanto observa a modificação das variáveis passo-a-passo no painel de ambiente, ainda, a cada execução se pode observar os erros e sucessos da execução acompanhando o console.

O software tem uma versão de licença livre e código aberto para contribuições, como também conta com versões profissionais de software e hospedagem com ônus financeiro⁷.

2.11.3 Software WEKA

Segundo o sítio de internet da Universidade de Waikato⁸, o WEKA é uma coleção de algoritmos aprendizado de máquina para tarefas de mineração de dados.

⁷ <https://www.rstudio.com/support/>.

⁸ <https://www.cs.waikato.ac.nz/~ml/weka/>.

Essa coleção inclui ferramentas para etapas de preparação, classificação, regressão, clusterização, associação de regras e visualização de dados. O software é de licença livre e de código aberto.

O WEKA possibilita a utilização de algoritmos de pré-processamento para seleção de atributos, permite a alimentação de dados a algoritmos de aprendizagem de máquina e analisar os resultados e as suas performances, tudo isso sem programar nenhuma linha (WITTEN; FRANK; HALL, 2011, p. 403)

A importação de dados desse software pode ser feita em arquivos com diversos tipos de extensão, porém a extensão padrão para uso no WEKA é a “. *arff*”. É possível adicionar um novo pacote de visualização de dados através do gerenciador de pacotes, esse possibilita a importação de arquivos diretamente do Excel ou de formato em CSV (WITTEN; FRANK; HALL, 2011).

Esse software é muito utilizado academicamente provavelmente por ter uma variedade muito grande de algoritmos de tarefas de KDD diferentes, possibilitar diversos pré-processamentos e transformações (como discretização, seleção de atributos, normalização, amostragem) sem ter a necessidade de conhecimentos de programação em Java, linguagem de programação usada na construção do WEKA.

Na próxima seção, no item Materiais e Métodos é explicado o uso de cada um desses softwares para cada etapa realizada. Além da descrição de cada etapa, a seção trata da caracterização da pesquisa, tabelas de pesquisa e outros assuntos dos encaminhamentos metodológicos.

3 ENCAMINHAMENTOS METODOLÓGICOS

Esta seção apresenta a caracterização da pesquisa, materiais e métodos e, análise de resultado.

3.1 CARACTERIZAÇÃO DE PESQUISA

A metodologia é a maneira utilizada pelo pesquisador para desenvolver uma pesquisa. Ao definir o termo “método”, Marconi e Lakatos (2003) oferecem a seguinte interpretação do termo:

o método é o conjunto de atividades sistemáticas e racionais que com maior segurança e economia permite ao pesquisador alcançar o objetivo. Neste sentido, através do método chega-se a conhecimentos válidos e verdadeiros, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista. (MARCONI; LAKATOS, 2003, p. 83).

Como a pesquisa visa verificar o perfil das vítimas de VD com resultado de morte e sendo esse um tema ainda pouco explorado, a abordagem é quantitativa (devido ao modelo de dados) e a tipologia utilizada é exploratória.

Segundo Freixo,

o método de investigação quantitativo tem por finalidade contribuir para o desenvolvimento e validação dos conhecimentos; oferece também a possibilidade de generalizar os resultados, de prever e de controlar os acontecimentos. (FREIXO, 2010, p. 145).

A lógica da aplicação de métodos e procedimentos do KDD nos dados dessa pesquisa é preditiva. Propõe-se generalizar o conjunto de dados a ponto de descrever seu comportamento e determinar um grupo de maior risco é fundamental para o sucesso nos resultados.

Gil (2008, p. 27) descreve as pesquisas exploratórias como aquelas que “têm como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores”.

Segundo Martins e Theóphilo (2009), para a condução de qualquer pesquisa científica, quando se pretende desenvolver um tema ou problema, é necessário ter por base as referências publicadas em livros, periódicos, sites, anais de congressos etc.

Essa pesquisa é orientada por meio de pesquisa bibliográfica, livros, artigos científicos, monografias, dissertações de mestrado, teses de doutorado com vistas a recolher, selecionar e interpretar as contribuições teóricas já existentes.

A abordagem quantitativa é a mais adequada devido ao modelo de dados que é disponibilizado pela SESP/PR.

QUADRO 7 – CLASSIFICAÇÃO DA PESQUISA

Abordagem	Quantitativa: os dados das bases de pesquisa são em sua maioria categóricos e analisados pela sua frequência. Essas bases são submetidas a métodos preditivos do KDD almejando descrever o comportamento desse grupo de vítimas para propor possíveis indivíduos que correm maior risco.
Método	Indutivo: o processo de Mineração de Dados proposto busca na repetição de características das vítimas e de ocorrências, encontrar regularidade e encontrar padrões desconhecidos.
Nível	Exploratória: tem a intenção de trazer um novo conhecimento científico e prático a fatos pouco estudados.
Envolvimento	Não Participante: o é, pois não existe o controle sobre os dados.
Bibliográfica	A pesquisa nas literaturas influentes da GI, VD e da KDD orientarão a pesquisa proposta.

FONTE: O Autor (2020).

3.2 MATERIAIS E MÉTODOS

Esta seção inicia descrevendo o modelo atual de recuperação de dados e produção de relatórios no CAPE, segue com as etapas do KDD: seleção, pré-processamento, transformação, mineração de dados e avaliação dos resultados.

3.2.1 Modelo atual de recuperação de dados e produção de relatórios no CAPE

O fluxo da informação é um fator importante para cada organização que gerencia a informação, como o CAPE. Explica-lo em um paralelo com o fluxo de Beal (2008, p.29), aproxima o conhecimento acadêmico da área da GI em prol descrever um processo que foi organizado para atender as necessidades da SESP/PR, necessidades estas que costumam mudar constantemente.

O software que armazena os dados mais importantes para a SESP/PR atualmente é o BOU. Então, no caso do BOU, a parte de Identificação de Necessidades é realizada pelo comitê gestor do BOU e os analistas da Companhia de Tecnologia da Informação e Comunicação do Paraná (CELEPAR); a Obtenção é

feita através do cadastro das polícias estaduais e dos cidadãos através do site da delegacia eletrônica (registra extravio de documentos e furtos desde 2017, estelionatos e violência doméstica desde 2021); o Tratamento é realizado pelo CAPE; o armazenamento é realizado pela CELEPAR; a Distribuição é realizada pelo CAPE gerenciando os sistemas da segurança pública e; o Uso é realizado pelas polícias do estado, a SESP/PR e o público não ligado à segurança pública que solicita dados e informações ao CAPE.

A recuperação de dados da SESP/PR feita pelo CAPE é organizada em periódica e por demanda. As periódicas ocorrem a cada três meses com um fechamento anual⁹, ambos publicados no site da SESP; as por demanda são designadas aos analistas e são solicitadas por público interno e externo. A extração de dados para cada área de especialidade é feita por pessoal especializado do próprio CAPE.

Dentre as recuperações periódicas, existem as que alimentam sistemas nacionais de dados de segurança pública, como o caso do sistema SINESP da Secretaria Nacional de Segurança Pública (SENASP) ou sistemas da própria SESP/PR (BI, CAPEGEO, e etc.).

Os analistas criminais e geocriminais do CAPE solicitam por demanda recuperações de dados específicos para produzirem seus relatórios, esses reportes geralmente tem no seu conteúdo números absolutos ou frequência relativa, uma explicação dos dados do relatório, a identificação do sistema usado para extração de dados e uma explicação sintetizada do comportamento dos dados.

Essas solicitações de dados são feitas principalmente pela própria SESP, outros órgãos do Estado do Paraná, SENASP, livre imprensa, pesquisadores acadêmicos e institutos de pesquisa.

No ano de 2021 foi disponibilizado o acesso a parte da ferramenta de *business intelligence* (BI) da CAPE ao público¹⁰. Os dados estão organizados em: crimes, mortes, violência, drogas e banco de dados. Dados de VD, violência contra mulher, entre outros, podem ser pesquisados acessando o ícone violência. No ícone mortes, pode-se pesquisar dados das mortes no Paraná, inclusive feminicídio separadamente.

⁹ <https://www.seguranca.pr.gov.br/CAPE/Estatisticas>

¹⁰ <https://bi2.pr.gov.br/single/?appid=058d7d0b-7d95-4ada-b6dd-81f50b47e814&sheet=273198fc-73f7-4347-aeb4-57f2d9966375>

Em base de dados, no organizador de conteúdo mortes violentas intencionais, é possível visualizar alguns dados de vários tipos de mortes intencionais, sendo possível selecionar separadamente os feminicídios de um ano.

3.2.2 Seleção de dados

Na etapa de KDD denominada **seleção**, utilizando a ferramenta PGAdmin (versão 4) foram selecionados via *script* SQL todos os atributos autorizados pelo CAPE dos boletins de vítimas mulheres, maiores de 18 anos, registrados no BOU de 2018, 2019 e 2020 que tenham a política pública “violência doméstica”.

Os dados de feminicídio foram extraídos separadamente no Sistema de Ocorrências Letais (SCOL)¹¹ usando como o filtro a natureza jurídica “feminicídio”. Detalhes dos atributos podem ser vistos no item 3.2.6 que trata das tabelas de pesquisa.

3.2.3 Pré-processamento de dados

No pré-processamento se iniciou um processo de verificação de integridade, ruídos e completude nas bases. Parte do resultado dessa primeira verificação foi descrito em detalhes para cada base de dados no item Tabelas de pesquisa.

Alguns atributos das duas bases selecionadas ofereciam a oportunidade de uni-las aumentando as possibilidades de análise. Um atributo comum nas duas bases, o atributo contendo o nome da vítima foi considerado como um possível identificador usado nessa união, para tanto, a avaliação de duplicidades nos nomes completos verificou casos de até 46 vítimas diferentes com o mesmo nome. Portanto, para tornar possível o uso do nome da vítima como identificador foi solicitada uma nova base de dados, essa contendo o nome da vítima concatenado ao nome da mãe da vítima. Através do PGAdmin (versão 4), foi implementado no CAPE um *script* em SQL que realizou a contagem de boletins e naturezas jurídicas do sistema BOU, agrupados por chave concatenada, respeitando os critérios da etapa de seleção.

¹¹ Sistema onde são cadastrados todos os detalhes de um homicídio. Como um homicídio gera uma investigação (causa possíveis, autor(es), circunstâncias) é comum que os dados mudem durante a investigação, portanto, as delegacias atualizam as informações de todos os homicídios nesse sistema gerido pela CAPE na SESP.

Utilizando a ferramenta Excel observou-se a frequência absoluta para cada atributo das bases BOU e SCOL. Esta ação objetivou observar o agrupamento dos dados por atributo e então, verificar quais atributos tem mais possibilidade de representar melhor os conjuntos de dados.

Os atributos que tiveram mais capacidade de representação dos conjuntos foram levados para análise estatística com mais detalhes, vistos nos itens que tratam dos resultados. Porém, alguns atributos muito longos ou que não estavam adequados para representação gráfica foram submetidos a tratamentos, esses são explicados na próxima etapa, transformação.

A base extraída do BOU que continha um atributo denominado chave, resultado da união do nome da vítima com o nome da mãe da vítima, essa continha a contagem de todos os boletins e das naturezas para cada vítima. Porém, esses dados não estavam todos alinhados para cada vítima. Para alinhar cada natureza jurídica com a sua respectiva vítima nessa etapa, utilizou-se a função *pivot_wider* do pacote *dplyr* no RStudio. Um exemplo do resultado dessa função pode ser visto no quadro abaixo.

QUADRO 8 – EXEMPLO DE LINHA COM DADOS AGRUPADOS POR VÍTIMA

Nome da Vítima	Boletins	Ameaça	Lesão Corporal	Injúria	Descumprimento de Medida Protetiva
Maria da Penha	7	3	2	2	1

FONTE: O Autor (2022).

Um fato importante sobre a base de dados é que cada linha da base, ou cada boletim, pode conter mais de uma natureza jurídica como ameaça e lesão corporal, no pré-processamento alinhar as informações cria várias possibilidades de análise de dados.

Após alinhar as naturezas às vítimas, através do RStudio utilizando o pacote *dplyr* foi importada a base de dados recém-criada. Depois, foram carregados os nomes das vítimas do SCOL em uma variável, comparou-se o conteúdo dessa com a base. Esse procedimento gera impossibilidade de exportação para arquivo se houver duplicidade na recuperação de nomes de vítimas. Checando o resultado, verificou-se 26 homônimas na comparação com os nomes do SCOL. No excel, foi adicionado ao nome dos homônimos um número sequencial para diferenciar a vítima fatal das outras vítimas.

Separadas vítimas fatais das não fatais da base, essa base foi carregada novamente no RStudio, assim, se repetiu o procedimento de busca na base comparando os nomes das vítimas com o conteúdo da variável como os nomes das vítimas do SCOL. No *script* que buscava os nomes foi adicionado um atributo denominado morte que no resultado positivo de comparação dos nomes, esse determinava o valor “S” para as vítimas encontradas e “N” para as demais condições. Tal procedimento possibilita a análise das naturezas das vítimas fatais e das não fatais separadamente, ou ainda, analisar como um todo.

A nova base, com o atributo morte continha 138.731 vítimas maiores de VD dessas, 144 continham o valor “S” para a morte. Os dados desta foram discretizados na etapa de transformação com o intervalo de classe igual a 2. O desbalanceamento da classe inviabilizava aplicar está base em algoritmos de classificação com o PRISM, CN2 e árvores de decisão. Optou-se por produzir uma amostra no formato “*undersample*”, todos as linhas dos valores “N” para morte teriam a mesma probabilidade de participar da nova amostra enquanto se conservou as 144 vítimas mortas. Os resultados tanto dos algoritmos quanto do conhecimento produzido a partir deles não poderia ser apoiado em dados replicados ou artificiais devido a seriedade com que se trata o objeto dessa pesquisa.

A seleção de atributos para mineração de dados ocorreu na base recém-criada com 20 atributos de 288 vítimas fatais e não fatais. Das naturezas observadas para as vítimas mortas, algumas não poderiam ser utilizadas na mineração como seis naturezas que descreviam o tipo da morte ou um local com morte e um tipo de lesão corporal, a contra autoridade ou agente. As naturezas que descrevem a morte poderiam criar viés nos algoritmos de mineração de dados e a lesão contra autoridade ou agente é um efeito colateral do atendimento a esse tipo de ocorrência, essa natureza não pode ser atribuída a vítima.

A seleção de atributos na base já limpa, discretizada e balanceada foi submetida a quatro técnicas de seleção de atributos com intuito de reduzir a dimensionalidade da base.

Alguns atributos são mais importantes que outros na mineração de dados, a seleção correta de atributos diminui redundâncias e ruídos que podem confundir os algoritmos (BORGES, 2006; SOUZA, 2017).

As técnicas de seleção de atributos CFS, Gain Rantio, Info Gain e ReliefF foram utilizadas na base de naturezas jurídicas das 288 vítimas. A comparação de

desempenho entre eles foi discutida nos resultados da mineração de dados.

3.2.4 Transformação dos dados

A transformação dos dados das três tabelas resultantes das etapas anteriores foi implementada através de *scripts* na linguagem de programação R através da ferramenta RStudio. Nessa etapa foram utilizados os pacotes da linguagem R: *readr*, *readxl*, *writexl*, *tidyverse*, *dplyr*, *lubridate*, *ggplot2* e *themis*.

Alguns atributos continham valores que dificultavam a leitura nas representações gráficas dos dados. Utilizando o pacote *dplyr* os termos foram modificados para facilitar a leitura na análise e organização dos dados.

Para a análise da base BOU foi necessário alterar alguns valores muito longos do atributo nomejuridico (natureza jurídica), simplificar os valores dos atributos diasemana e período.

Os valores que foram modificados podem ser vistos em detalhes no APÊNDICE 1. Como exemplo, seguem abaixo as primeiras naturezas jurídicas simplificadas.

QUADRO 9 - SIMPLIFICAÇÃO DE VALORES NO ATRIBUTO NATUREZA JURÍDICA DO BOU

Original	DESCUMPRIR DECISAO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS DE URGENCIA PREVISTAS NESTA LEI
Simplificado	DESCUMBRIR ORDEM JUDICIAL DE MEDIDA PROTETIVA
Original	LESAO CORPORAL - SE A LESAO FOR PRATICADA CONTRA AUTORIDADE OU AGENTE DESCRITO NOS ARTS. 142 E 144
Simplificado	LESAO CORPORAL CONTRA AUTORIDADE OU AGENTE
Original	ESTA LEI CRIA MECANISMOS PARA COIBIR E PREVENIR A VIOLENCIA DOMÉSTICA E FAMILIAR CONTRA A MULHER, NOS TERMOS DO 80 DO ART. 226 DA CONSTITUICAO FEDERAL
Simplificado	VIOLENCIA CONTRA MULHER MARIA DA PENHA

FONTE: O Autor (2022).

Também os valores dos atributos do BOU diasemana e periodo foram modificados. Os valores substituídos são vistos nos próximos quadros.

Os dias da semana foram extraídos com um número no início para facilitar a indexação dos valores, porém, acaba imobilizando a possibilidade de indexar os valores de outra forma. Os termos simplificados facilitam a leitura, organização e possibilitar indexar o conjunto de dados de outras maneiras.

QUADRO 10 – MODIFICAÇÃO DOS VALORES DO ATRIBUTO DIASEMANA DO BOU

Original	Simplificado
0_Domingo	Dom
1_Segunda-feira	Seg
2_Terça-feira	Ter
3_Quarta-feira	Qua
4_Quinta-feira	Qui
5_Sexta-feira	Sex
6_Sabado	Sab

FONTE: O Autor (2022)

Ocorreu o mesmo com o atributo período, os valores começando com um número criavam um ordenamento único. O valor madrugada foi diminuído para facilitar a leitura e representação dos dados.

QUADRO 11 - MODIFICAÇÃO DOS VALORES DO ATRIBUTO PERIODO DO BOU

Original	Simplificado
4_Noite	Noite
3_Tarde	Tarde
2_Manha	Manha
1_Madrugada	Madrug

FONTE: O Autor (2022).

Na tabela de dados do SCOL continha um atributo idêntico para dia da semana, portanto, os valores foram simplificados a exemplo da tabela extraída do BOU.

QUADRO 12 - MODIFICAÇÃO DOS VALORES DO ATRIBUTO DIA DA SEMANA DO SCOL

Original	Simplificado
7-DOM	Dom
1-SEG	Seg
2-TER	Ter
3-QUA	Qua
4-QUI	Qui
5-SEX	Sex
6-SAB	Sab

FONTE: O Autor (2022).

Ainda, a base de dados de naturezas agrupadas por vítima foi discretizada com intervalo “2” para classe. Os atributos numéricos continham valores inteiros com

o valor mínimo 0 e máximo 13 com exceção dos atributos boletins que o valor mínimo é 1.

3.2.5 Mineração de dados

Para etapa de mineração de dados, verificou-se que devido às características das três bases estudadas, somente a base de ocorrências letais e a base de naturezas podem ser levadas aos algoritmos de indução de regras. Então, essas bases de dados foram analisadas sob a ótica da estatística descritiva e de algoritmos indução de regras de classificação: PRISM e CN2.

O PRISM utiliza a entropia como método de avaliação interna do algoritmo, o CN2 tem três opções para essa avaliação interna sendo a entropia uma delas. Então, a opção pela entropia no CN2 é importante quando o objetivo é comparar resultados. Para o uso desses algoritmos é importante ressaltar que o CN2 está implementado no Orange e o PRISM no WEKA.

As ferramentas utilizadas nessa etapa com exceção do Excel, são gratuitas, muitas pesquisas ocorrem sem fomento ou incentivos financeiros, ou ainda, o apoio financeiro não possibilita a aquisição de licenças de ferramentas utilizadas na etapa de mineração de dados. O uso de ferramentas licenciadas poderia comprometer a replicabilidade dessa pesquisa, portanto, as ferramentas RStudio, WEKA e Orange, de uso público, podem viabilizar novas pesquisas com dados de VD de outras fontes, nos mesmos moldes dessa dissertação.

3.3 TABELAS DA PESQUISA

As tabelas utilizadas nas análises são: vítima (SCOL), boletim (BOU), envolvido (BOU) e naturezas (BOU) sendo a base BOU resultado de junção de dados da tabela boletim e envolvido.

3.3.1 Base SCOL

A base de dados SCOL aqui tratada contém 237 linhas de dados somente de feminicídio. Cada linha contém dados de uma vítima e os demais dados do feminicídio relacionados a esta, exceto por uma vítima que constava em duas linhas, essas

cadastradas por delegacias diferentes, portanto, apesar do número de linhas existem 236 vítimas nessa base.

Como um exemplo de como uma vítima pode receber dois registros, foi relatado em 2020 no CAPE uma morte ocorrida em um município de responsabilidade de uma delegacia, na investigação foi constatado que o fato ocorreu na região de responsabilidade de outra delegacia. Tal fato gerou o registro da mesma vítima por duas delegacias. Segundo o analista responsável por verificar os homicídios, é realizada uma verificação minuciosa nos casos de homicídio registrados no SCOL, um processo de homologação de dados antes das publicações oficiais do estado é realizado periodicamente pelos analistas e estatísticos.

Essa é uma importante fonte de dados para os relatórios produzidos na CAPE sobre os crimes de feminicídio. A tabela extraída do sistema SCOL será tratada nessa dissertação como “base SCOL”.

Esta base é protegida por lei federal, lei estadual e decreto estadual, sendo assim, nessa pesquisa somente poderá ser tratada por este pesquisador e a orientadora da pesquisa, conforme termo de cooperação técnica. Por sua vez, durante o tratamento, haverá a anonimização dos dados sensíveis por meio da substituição destes por identificadores e a base resultante, por sua vez, será depositada na curadoria de dados sobre violência doméstica da UFPR.

Um exemplo de atributo que será trocado por identificador numérico é o caso do “nome”, pois esse somente é visto ou tratado pela equipe responsável por essas informações na SESP.

Além das substituições já citadas, alguns atributos dessa base serão suprimidos antes da entrega à curadoria por questões referentes à proteção legal dessa. Os campos: “unidade”, “logradouro”, “numero”, “controle”, “solucionado”, “apelido”, “codnatureza”, “dt_registro”, “grauparentesco”, “dt_alterado”, “parenteenvolvido” e “autoria_conhecida” não integrarão a base ao término desse trabalho de pesquisa.

No caso da tabela do SCOL, todos os campos extraídos pertencem a mesma tabela, a de vítimas cadastradas no sistema. Portanto, não houve necessidade de juntar dados de tabelas diferentes. Todos os campos dessa base podem ser vistos em detalhes no QUADRO 13 abaixo.

QUADRO 13 – ATRIBUTOS DA BASE SCOL

Coluna	Tipo de Dado	Descrição	Exemplo
anoboletim	numérico	Ano do boletim de ocorrência. O boletim pode ser registrado em delegacia de polícia civil até 180 dias da data do crime. Este ano se refere ao registro no sistema.	2020
codboletim	numérico	Número do boletim de ocorrência	165416
datafato	data	Data do início do fato (exata ou estimada pelo noticiante)	10/01/2020
ano	numérico	Ano do fato	2020
mes	numérico	Mês do fato	1
dia	numérico	Dia do mês do fato	11
diasemana	caractere	Dia da semana do fato	6-SAB
cidade	caractere	Município onde ocorreu o fato	ABATIA
bairro	caractere	Bairro onde ocorreu o fato	CENTRO
logradouro	caractere	Logradouro onde ocorreu o fato	Rua Mariano Torres
numero	numérico	Número do Logradouro onde ocorreu o fato	13
natureza	caractere	Natureza Jurídica do fato na investigação	FEMINICIDIO
vitima	caractere	Nome da vítima	Vedado
autoria_conhecida	caractere	Autor do fato foi identificado	S / N
motivacao	caractere	No início do processo a morte ocorreu por motivo de qual ação	DESENTENDIMENTO
idade	numérico	Idade da vítima	25
raca_cor	caractere	Cor da pele da vítima	AMARELA/BRANCA/PARDA/PRETO/NAO INFORMADO
sexo	caractere	Sexo da vítima	Todos os indivíduos da base são (F) nesse campo
dt_nascimento	data	Data de nascimento da vítima	04/05/1974
localresidencia	caractere	Município de residência atual da vítima	CURITIBA
estadocivil	caractere	Estado do civil da envolvida/vítima	CASADA/CONVIVENTE/SOLTEIRA/SEPARADA/DIVORCIADA/MUVA/O UTROS/NAO INFORMADO
idade_tipo	caractere	Unidade de medida do campo idade-2	Anos/Meses/Dias
nacionalidade	caractere	Nacionalidade da vítima	Brasileira
orientacao_sexual	caractere	Orientação sexual da vítima	HETERO/HOMOSEXUAL
profissao	caractere	Profissão da vítima	DONA DE CASA/VENDEDORA/EMPRESARIA/DESEMPREGADA/ESTUDANTE ...
meioempregado	caractere	Meio empregado na ação	AGRESSAO/FACA/REVOLVER ...
grauinstrucao	caractere	Grau de instrução da vítima	NAO ALFABETIZADO/PRIMEIRO GRAU INCOMPLETO/PRIMEIRO GRAU ... até TERCEIRO GRAU COMPLETO
identidadegenero	caractere	Identidade de gênero da vítima apurado na investigação (Campo aberto). Último Campo adicionado no sistema SCOL	Em 10 linhas consta (Cisgênero), demais linhas estão nulas ou vazias.

FONTE: O Autor (2021).

O atributo “idade_tipo” demonstrado acima possui três possibilidades (Anos, Meses e Dias), porém foi convencionado o uso de anos completos para esse cadastro, ou seja, se um indivíduo nessa base tiver 20 anos e 4 meses e 15 dias, será registrado “20” na “idade” e “Anos” na “idade_tipo”.

O atributo “natureza” no QUADRO pode sofrer alteração no conteúdo quando na investigação do crime ou o poder judiciário (ministério público) reenquadra a natureza jurídica. O arquivo utilizado nessa dissertação contém somente registros de feminicídio que já foram enquadrados dessa forma até a extração em outubro de 2021.

Essa tabela também possui atributos idênticos em formato para cadastro durante as apurações das investigações. Os atributos “dt_nascimento-2”, “idade-2”, “raca_cor-2”, “sexo-2” e “codnatureza-2” são usados para cadastro durante e depois da investigação, portanto podem apresentar valores diferentes dos atributos do cadastro inicial. Esses não estão no QUADRO 13 por serem exatamente iguais aos atributos do cadastro inicial, o que causaria repetições desnecessárias.

O atributo “motivacao” possui mais dois atributos idênticos para serem usados no curso da investigação, “motivacao-2” e “motivacao-3”. Os referidos atributos armazenam dados como os da “motivacao”, mas oferecem a possibilidade de acrescentar detalhes a cada vítima. A ideia de ter mais de um campo “motivação” surgiu da necessidade de confrontar dados iniciais e da investigação. Também, nos boletins cadastrados no BOU é possível inserir mais de uma motivação, o que tornaria os sistemas com cadastro mais parecido.

A motivação é o único caso de atributo nessa base que tem um terceiro item de cadastro para representá-lo em tempos diferentes, “motivacao-3”. Segundo o analista de mortes da CAPE, o atributo foi criado quando alguns usuários do sistema solicitaram.

Eventualmente pode-se suprimir um ou alguns atributos na análise estatística que tenham informações repetidas, como o caso da “data/hora” que foi extraída também em partes: “dia”, “mês”, “ano” e “diasemana”. Tratar o campo separadamente pode enriquecer os resultados.

3.3.2 Base BOU

A base BOU contém 368.476 linhas, composta por dados de ocorrências criminais com a política pública “VIOLENCIA DOMESTICA” e que tenha pelo menos uma vítima mulher de idade maior ou igual a 18 anos. A política pública nesse sistema informatizado é um atributo que auxilia no agrupamento dos dados para a estatística oficial do estado.

Na observação dos atributos da base constatou-se que “curtis” e “estadocivil” possuem 49 linhas nulas, “grau_instrucao” 72 linhas nulas e “ocupacao” 132.583 linhas nulas. Em menor valor, “cidade” e “bairro” possuem linhas nulas em 4 e 8 ocasiões respectivamente.

Avaliando valores vazios para colunas foi observado a ausência de 7.550 dados em “grupo_meio_empregado”, “meio_empregado” e “nomejuridico”.

O conteúdo do sistema BOU utilizado nesse estudo é composto por uma pequena parte importante dos dados deste sistema. Foi autorizado pelo CAPE a extração de alguns atributos de duas tabelas. A tabela resultado dessa extração será tratada nessa dissertação como “base BOU”.

Os atributos dessa base podem ser vistos com maiores detalhes no QUADRO 14 abaixo.

QUADRO 14 – ATRIBUTOS DA BASE BOU

Coluna	Tipo de Dado	Descrição	Exemplo
base	caractere	Identificação da base. Exemplo: Autores e vítimas extraídas separadamente e unificadas com comando 'union'.	VD_Mulher_Maior
id	numérico	Chave com número do boletim de ocorrência concatenado com o ano do boletim	2020325145
anoboletim	numérico	Ano do boletim de ocorrência. O boletim pode ser registrado em delegacia de polícia civil até 180 dias da data do crime. Este ano se refere ao registro no sistema.	2020
codboletim	numérico	Número do boletim de ocorrência	165416
fatoinicial	data	Data do início do fato (exata ou estimada pelo noticiante)	10/01/2020
ano	numérico	Ano do fato	2020
mês	numérico	Mês do fato	1
mês_extenso	caractere	Mês do fato por extenso	1_janeiro
dia_semana	caractere	Dia da semana do fato	3_Quarta-feira
período	caractere	Período do dia que ocorreu o fato	Manha
hora	numérico	Hora do fato	2_Manha
uf	caractere	Estado do Fato. O bou recebe fato ocorridos no estado do Paraná.	PR
cidade	caractere	Município onde ocorreu o fato	ABATIA
bairro	caractere	Bairro onde ocorreu o fato	CENTRO
nome	caractere	Nome da vítima	Vedado o exemplo
idadeestimada	numérico	Idade estimada ou coletada com com a envolvida vítima	32
curtis	caractere	Cor da pele informada pela envolvida/vítima	Branca/Parda/Preta/Amarela/Nao Informado
estadocivil	caractere	Estado do civil da envolvida/vítima	Casada/Convivente/Solteira/Separada/Divorciada/Viúva/Outros/Nao informado
grau_instrucao	caractere	Grau de instrução da envolvida/vítima	Nao alfabetizado/Primeiro grau incompleto/Primeiro grau completo/Segundo grau incompleto/ Segundo grau completo/Terceiro grau incompleto/Terceiro grau completo/Nao informado
ocupacao	caractere	Ocupacao (Trabalho formal ou informal) da envolvida/vítima	DONA DE CASA/VENDEDORA/EMPRESARIA/DESEMPREGADA/ESTUDANTE ...
grupo_meio_empregado	caractere	Grupo de meios empregados da tabela bou	PESSOAS/ELETRONICOS/ARMAS BRANCA/ARMAS DE FOGO/FERRAMENTAS ...
meio_empregado	caractere	meio empregado no ação delituosa	AGRESSAO/FACA/REVOLVER ...
nomejuridico	caractere	Nome do crime ou contravenção segundo o ordenamento legal atual. Foram filtradas somente ocorrências criminais	Ameaça / Lesão Corporal

FONTE: O Autor (2021).

Importante ressaltar que a base de dados apresentada no QUADRO 14, atributos da base BOU, é resultado de uma extração unindo informações da tabela “boletim” e da tabela “envolvido” do banco de dados do sistema BOU. Os atributos “id”, “anoboletim”, “codboletim”, “fatoinical”, “uf”, “cidade”, “bairro” e “nomejuridico” são extraídos diretamente da tabela “boletim”. Os atributos “ano”, “mes”, “mes_extenso”, “dia_semana”, “hora” e “periodo” são calculados através do campo “fatoinicial”, ou seja, são pré-processados durante a extração. Os atributos “nome_vitima”, “codenvolvido”, “idadeestimada”, “curtis”, “estadocivil” e “grau_instrucao” pertencem a tabela envolvido. As tabelas são unificadas na etapa de extração através do campo id (número do boletim concatenado com ano) através do comando “JOIN”.

3.3.3 Base Naturezas

A Base Naturezas foi extraída do sistema BOU agrupando todas as naturezas jurídicas das ocorrências por vítima. O estudo foi idealizado inicialmente com a base SCOL e base BOU porém na tentativa de agrupar vítimas pelas naturezas surgiu um problema relacionado à vítimas com o mesmo nome. Foi considerado importante o agrupamento das naturezas por vítima tanto para a estatística descritiva quanto para o processo de mineração de dados. Portanto, foi solicitado ao CAPE uma base pré-processada com a contagem de naturezas agrupada por uma chave concatenada composta pelo nome da vítima e o nome da mãe, diminuindo a possibilidade de existir homônimos, o que causaria uma contagem que não condiz com a realidade.

QUADRO 15 – ATRIBUTOS DA BASE NATUREZAS

Coluna	Tipo de Dado	Descrição	Exemplo
chave	caractere	É composto pelo nome da vítima concatenada com o nome da mãe da vítima.	JOSE MARIA DA SILVA-JOSEFA DA SILVA
ocorrencia	caractere	Natureza jurídica do BOU ou botelins registrados agrupados por chave (vítima+mãe)	AMEACA/boletins/LESAO CORPORAL ...
count	numérico	Contagem das naturezas e boletins de ocorrência agrupados por vítima.	1

FONTE: O Autor (2021).

Esta base é composta por três tributos, vistos no QUADRO 15. O segundo atributo da base, “ocorrencia”, recebe vários valores de naturezas jurídicas e um valor

“boletins” para todas as vítimas. A contagem agrupada é feita para todas as vítimas de violência doméstica cadastrada no sistema BOU nos anos de 2018, 2019 e 2020.

Apesar da extração com pré-processamento das naturezas jurídicas agrupada por vítima, ainda foi necessário o pré-processamento da base no RStudio antes de exportar a base Naturezas usada nas análises.

3.4 AVALIAÇÃO DO MODELO PREDITIVO

Ao se avaliar um método preditivo recomenda-se pela literatura especializada, incluindo as já citadas na seção 2, a utilização de métricas próprias desses métodos. Dentre elas a análise de matriz de confusão, acurácia, sensibilidade, especificidade, precisão e F-Score.

As matrizes de confusão e as medidas resultantes dos métodos são comparadas e discutidas nos resultados e considerações finais. Portanto, a capacidade preditiva dos algoritmos é analisada observando todo o conjunto de resultados, não somente a acurácia.

Ainda, a literatura sugere que ao se avaliar um modelo preditivo deve-se acrescentar a análise qualitativa aos resultados, portanto, os resultados foram avaliados por dois especialistas do CAPE.

Segundo Halmenschlager (2002, p.60),

além de uma avaliação objetiva, é necessária uma avaliação qualitativa dos resultados encontrados, por parte do especialista. É possível que, mesmo com níveis de precisão aceitáveis, o especialista interprete os resultados da mineração da seguinte forma: a) fica satisfeito com os resultados e surpreso com alguns dos padrões obtidos; b) fica satisfeito com os resultados, mas percebe que já conhecia os padrões obtidos; c) fica insatisfeito com os resultados. (apud JOHN, 1997¹²).

De fato, por mais que se aprove um modelo preditivo através das medidas e análise de matriz de confusão, como esse modelo de classificação, a avaliação subjetiva do especialista é importante nos quesitos: utilidade, aplicabilidade e possibilidade de criar um plano de contingência com os resultados aprovados dentre outros.

¹² JOHN, George H. **Enhancements to the Data Mining process**. 1997. Tese (Doctor of Philosophy) – Department of Computer Science, Stanford University, Stanford.

4 RESULTADOS E ANÁLISES

Esta seção trata de resultados da estatística descritiva da base BOU, base SCOL, SCOL *versus* BOU, base Naturezas *versus* feminicídio e resultados da aplicação do KDD nas bases Naturezas e SCOL.

4.1 RESULTADOS DA ESTATÍSTICA DESCRITIVA

Na observação das bases de dados utilizando contagem de atributos agrupada foi possível identificar um fator tratado na seção 2, a revitimização em que se observou 23.680 vítimas com mais de um boletim registrado nos três anos estudados. Ainda, foi verificado que 58 vítimas registraram entre 10 e 19 boletins no mesmo período.

4.1.1 Estatística descritiva da base SCOL

A base SCOL contém 236 vítimas de feminicídio para os três anos, 68 em 2018, 93 em 2019 e 75 em 2020. Do ano de 2018 para 2019 observou-se um crescimento de 37% nos feminicídios e de 2019 para 2020 observou-se uma queda de 19%.

A idade média das vítimas da base SCOL é de aproximadamente 37 anos nos três anos estudados. Em 2018, 36,51 anos, em 2019, 37,39 anos e em 2020, 37,06 anos, discriminados em maiores detalhes na TABELA 2. O coeficiente de variação maior que 32,2% para os três anos estudados indica que as idades nessa base têm alta dispersão, ou seja, é uma distribuição heterogênea. A alta dispersão observada não contribui para aceitar a idade média como a melhor forma de generalizar esse grupo de vítimas.

Valores menores de 18 anos foram removidos (9 linhas) da estatística descritiva apresentada na TABELA 2.

TABELA 2 - ESTATÍSTICA DESCRITIVA DO ATRIBUTO IDADE DA BASE SCOL

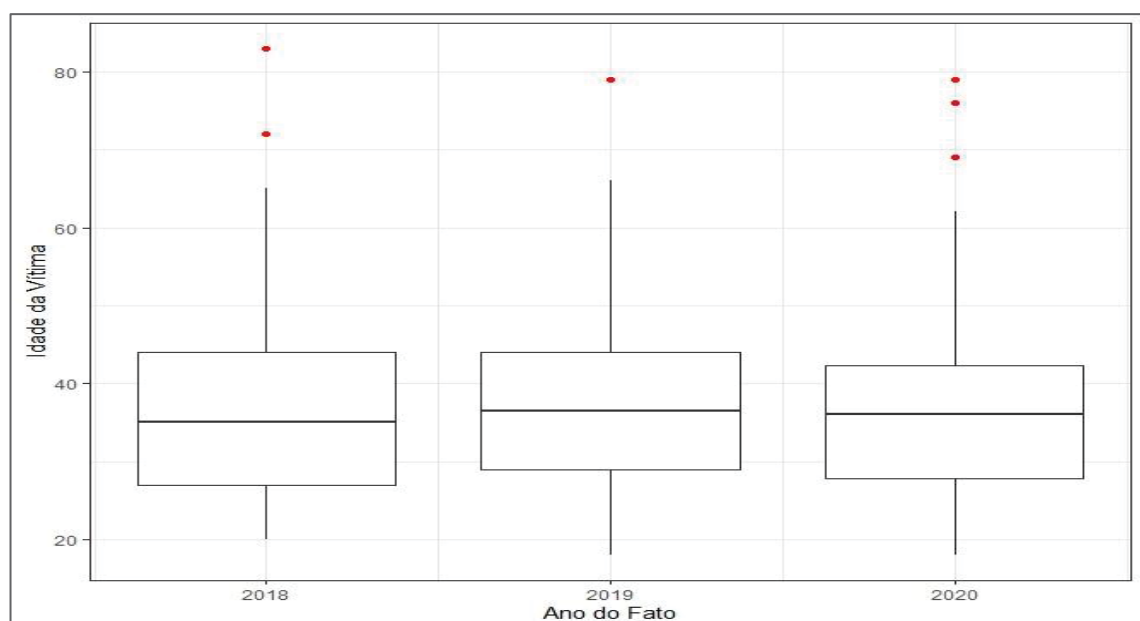
Ano	Média	Desvio Padrão	Coefficiente de Variação (%)	1º Quantil	3º Quantil
2018	36,51	12,77	34,98	27	44
2019	37,39	12,05	32,25	29	44
2020	37,06	12,87	34,74	28	42

FONTE: O Autor (2021).

Dentre as idades observadas, 6 linhas continham valor abaixo dos 17 anos, são considerados valores atípicos. Das 9 vítimas fatais dessa base com idade menor de 18 anos, uma teve seu registro no ano de 2018, cinco no ano de 2019 e três no ano de 2020. O valor máximo observado nesse atributo foi 83 anos e o mínimo 0. Cullen; Fritzson (2019) observaram que em uma amostra de 141 eventos com 416 mortes por violência doméstica intrafamiliar ocorridas na Austrália mais de 90% tiveram filhos como alvos. A presença de indivíduos de menos de 17 anos cadastrada no sistema de ocorrências letais tem grandes chances de ter ocorrido no ataque as suas mães, vítimas de feminicídio.

Os *outliers* representados em vermelho no GRÁFICO 1 são valores muito acima da média.

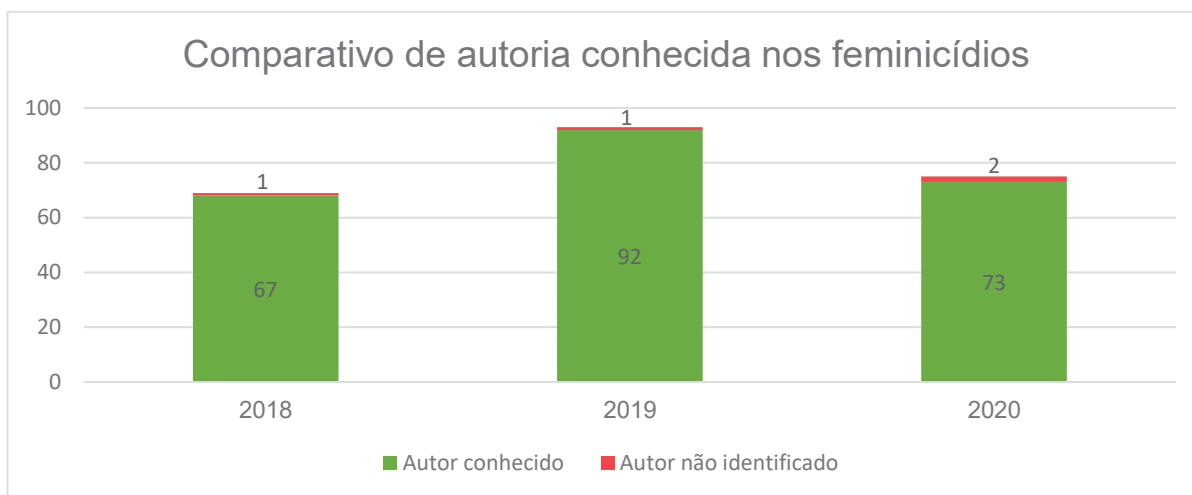
GRÁFICO 1 – COMPARATIVO DAS IDADES DAS VÍTIMAS DE FEMINICÍDIO DA BASE SCOL AGRUPADA POR ANO



FONTE: O Autor (2022).

Uma das características dos crimes de feminicídio no Paraná identificada na base do SCOL, é que a maioria desses tem autor conhecido. No ano de 2020, aproximadamente 97% dos registros de feminicídio tinham autor conhecido, em 2019 e 2018, 99%.

GRÁFICO 2 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DOS AUTORES CONHECIDOS EM FEMINICÍDIOS DA BASE SCOL AGRUPADOS POR ANO



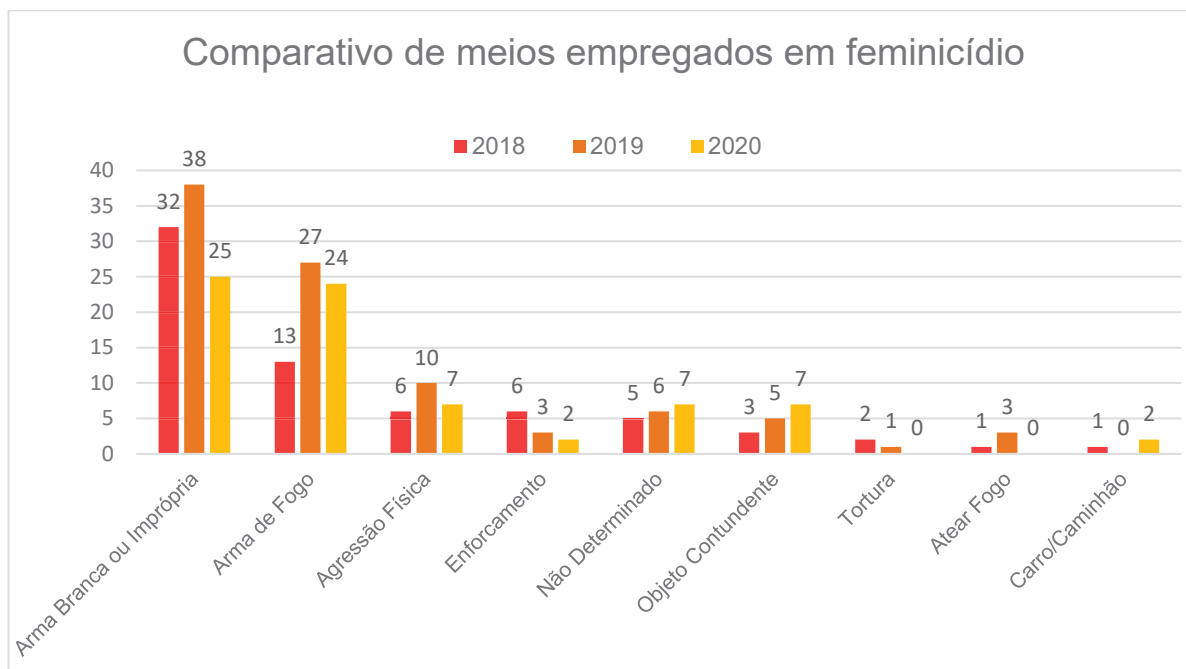
FONTE: O Autor (2022).

As motivações Passional e Violência doméstica representam aproximadamente 83% em todos os crimes de feminicídio.

A base SCOL contém o atributo “meio_empregado”, referente aos meios empregados no crime. Somados nos três anos 77% das mortes tiveram os meios empregados: “Arma branca ou imprópria” 40%; “Arma de fogo” 27% e; “Agressão física” 9,7%.

O item Carro/Caminhão é resultado da soma de um atropelamento por caminhão e dois atropelamentos por carros. Foi tratado com o analista de ocorrências fatais do CAPE sobre a presença desse tipo de ocorrência. Foi relatado pelo analista que, um caso registrado como homicídio culposo no trânsito durante as investigações pode ser modificado após novas provas e informações, modificando até a natureza jurídica. Portanto, um homicídio culposo no trânsito pode se tornar homicídio doloso ou feminicídio, fato que pode ter acontecido em algumas dessas três mortes. O GRÁFICO 3 representa esses meios agrupados por ano.

GRÁFICO 3 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DOS MEIOS EMPREGADOS NOS FEMINICÍDIOS DA BASE SCOL AGRUPADOS POR ANO



FONTE: O Autor (2022).

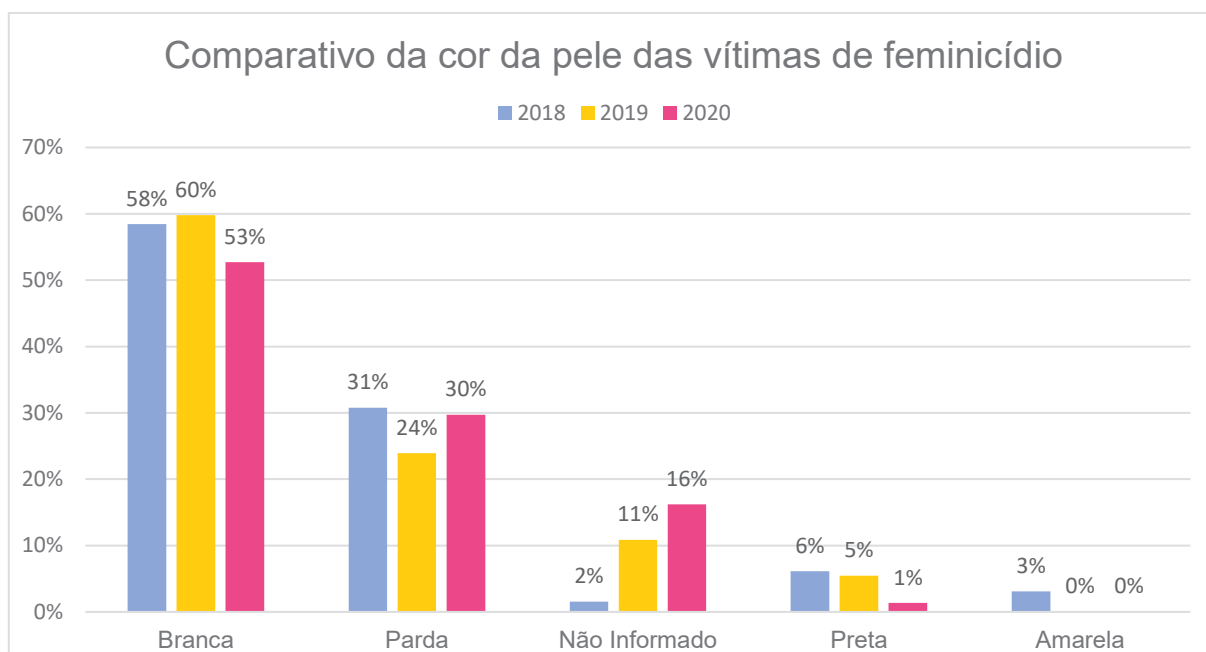
Observa-se uma oscilação primeiramente crescente e depois decrescente quanto ao uso de Arma branca ou imprópria, uso de Arma de fogo, e Agressão física. Nota-se um decréscimo gradual do Enforcamento, o crescimento ordenado do Não determinado e do uso de Objeto contundente, queda ordenada da Tortura, oscilação de Atear fogo e um crescimento dos Atropelamentos no último ano, 2020. Com exceção de Enforcamento, Não determinado, Objeto contundente, Tortura e Atropelamento.

O atributo “raça/cor” continha dados em 232 linhas, restando 3 sem dados. Na frequência relativa dos três anos raça ou cor amarela representa em 1% dos casos, já a branca aproximadamente 60%, seguida de parda 28%, preta 4% das vítimas dessa base de dados.

Segundo censo de 2010 feito pelo Instituto Brasileiro de Geografia e Estatística (IBGE) a população do Paraná branca representa 70,1%, preta 3,1%, parda 25,4%, amarela 1,2%, indígena 0,2%. A distribuição de “raça/cor” da tabela tem percentuais parecidos com a distribuição de população publicada pelo IBGE com exceção da cor parda e branca. Das mortes com natureza feminicídio no SCOL as vítimas brancas são quase 10% a menos da proporção dessa cor da pele no estado,

já a parda tem 2,6% a mais comparados a mesma proporção. O GRÁFICO 4 demonstra a proporção de vítimas pelo atributo “raça/cor”.

GRÁFICO 4 – COMPARATIVO DA FREQUÊNCIA RELATIVA DA COR DA PELE DAS VÍTIMAS DE FEMINICÍDIO DO SCOL

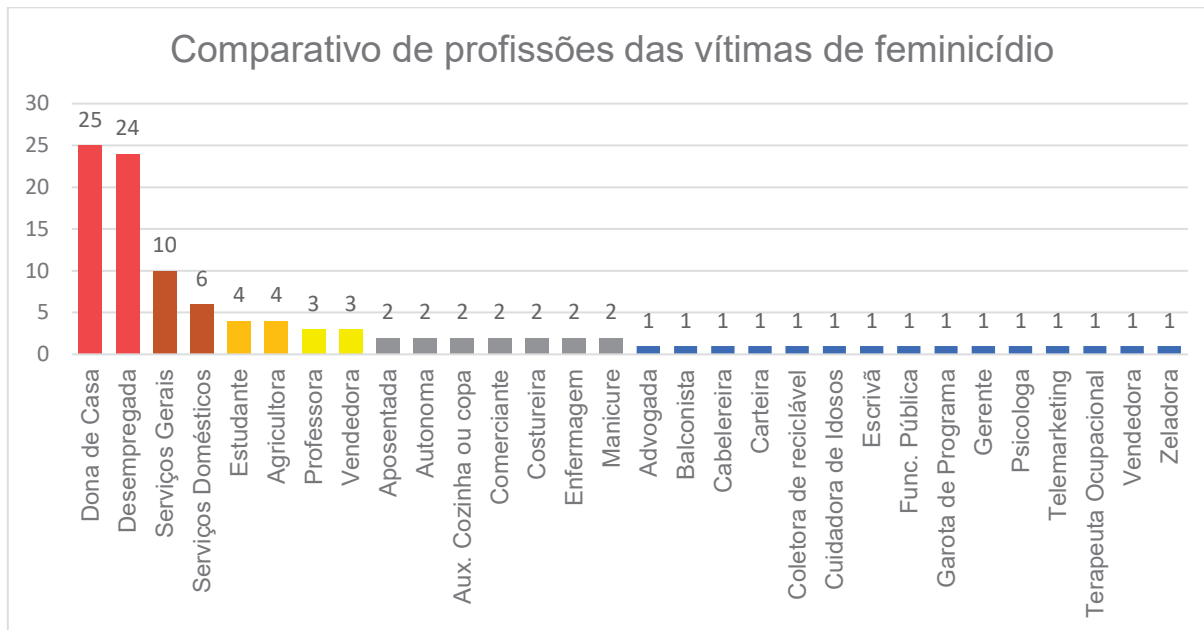


FONTE: O Autor (2022).

Nas proporções representadas neste gráfico é possível notar uma diminuição da cor branca no último ano e o inverso no caso da cor parda. O crescimento do não informado é um fato que diminui a capacidade da representação das vítimas por cor da pele. As vítimas de cor preta tiveram uma diminuição bem aguda no último ano, mesmo ano que o não informado teve incremento de 5% sobre o ano anterior. Todos os casos registrados com amarela ocorreram em 2018.

Na base SCOL também é possível registrar a profissão da vítima. Este atributo, “profissao”, em muitos casos não foi preenchido pela delegacia devido a não se ter confirmação desse dado. Na base constavam 108 registros preenchidos com valores válidos, esses são demonstrados no GRÁFICO 5. Os valores considerados inválidos para a representação gráfica representam 54% desse atributo.

GRÁFICO 5 – COMPARATIVO DE FREQUÊNCIA ABSOLUTA DAS PROFISSÕES DAS VÍTIMAS DE FEMINICÍDIO NO SCOL



FONTE: O Autor (2022).

Alguns itens do GRÁFICO 5 foram unificados. O item Dona de Casa é composto pela soma de dona de casa e do lar, Serviços Gerais é a soma de auxiliar de serviços gerais e serviços gerais, Serviços Domésticos é soma de empregada doméstica e diarista, Agricultora é a soma de agricultora e trabalhadora rural, Professora é a soma de professora e professora municipal, Enfermagem é a soma de técnica em enfermagem e enfermeira.

A escolha por representar o gráfico de profissões agrupados nos três anos se deu pois era necessário representar todas as profissões que foram cadastradas nos feminicídios. O número de linhas da base e a quantidade de profissões diferentes registradas são melhor representadas dessa maneira.

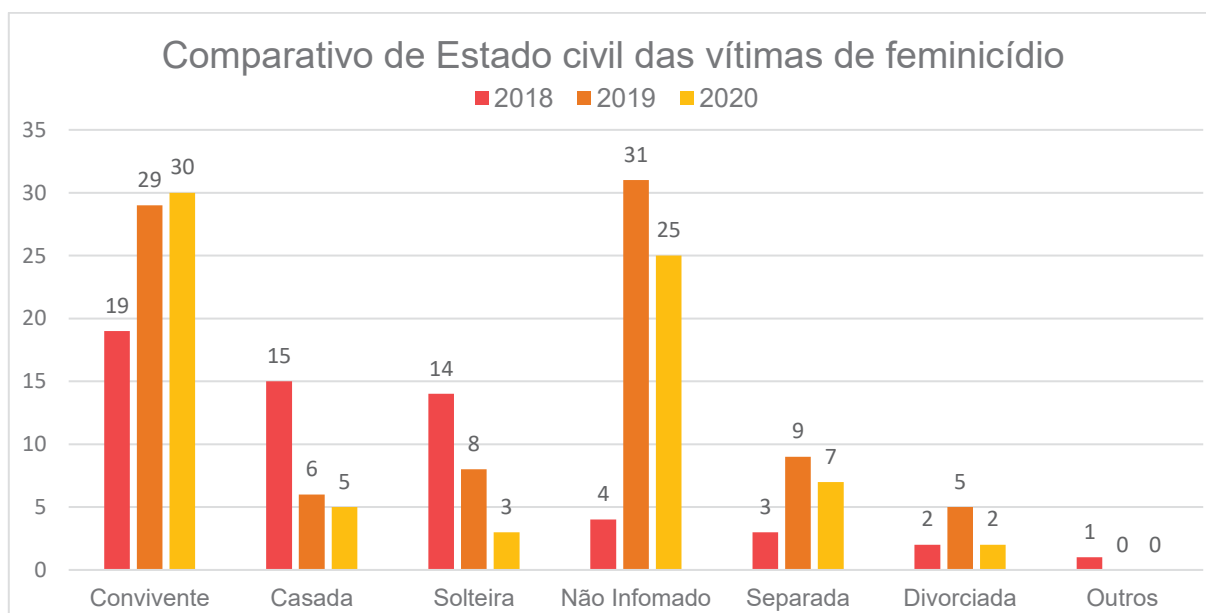
Observado na base que o atributo que representa a Orientação Sexual tem problema referente a completude, como a profissão, já discutida. Na base SCOL essa coluna continha 156 linhas preenchidas, sendo 155 Heterossexuais e 1 Homossexual. Uma maior completude oportunizaria uma discussão mais aprofundada.

O atributo que armazena a Identidade de gênero foi adicionado no fim de 2019 e tem muitas linhas sem preenchimento. Foi verificado que 10 linhas foram preenchidas, todos eram Cisgênero. Segundo o dicionário de língua portuguesa

Priberam, cisgênero é atribuído a quem “tem uma identidade de gênero idêntica ao sexo que foi atribuído à nascença”.

O atributo “estado_civil” das vítimas de feminicídio continha 218 linhas preenchidas. Na frequência relativa dos três anos o estado civil “Convivente” representa aproximadamente 36% dos casos dessa base. Outros valores, “Casada” e “Solteira” representaram aproximadamente 12%, “Separada” aproximadamente 8%, “Divorciada” 4%. Ainda, “Não informado” representa aproximadamente 28%.

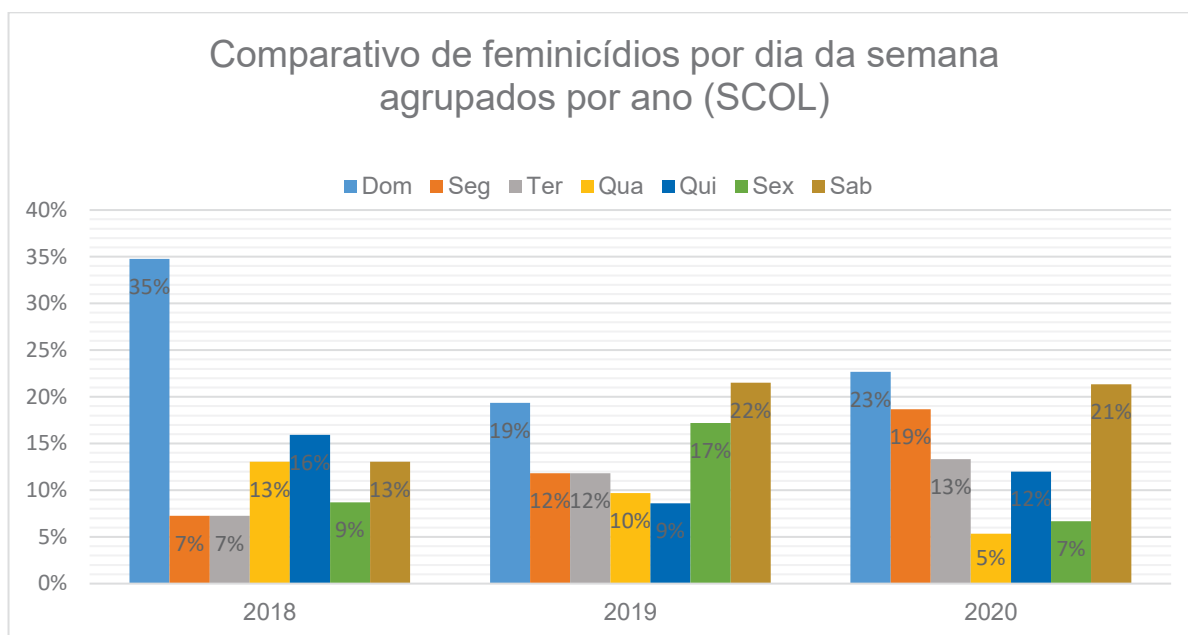
GRÁFICO 6 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO ESTADO CIVIL DAS VÍTIMAS DE FEMINICÍDIO DO SCOL AGRUPADOS POR ANO



FONTE: O Autor (2022).

O estado civil “Convivente” teve um comportamento de crescimento no período, destacando-se o incremento de 10 casos em 2019. “Casada” e “Solteira” tiveram comportamento decrescente, sendo o caso de “Solteira” um decréscimo ordenado, diminuíram cinco casos ao ano. O crescimento do “Não informada” nos últimos anos é um fato preocupante para representar os dados, como no gráfico anterior. “Separada” cresceu 200% de 2018 para 2019, e decresceu 22% de 2019 para 2020, sendo um comportamento bem diferente dos outros valores. “Divorciada” também oscilou, mas teve um crescimento de 150% de 2018 para 2019, quando que “Outros” permaneceu nulo nos últimos dois anos.

GRÁFICO 7 - COMPARATIVO DA FREQUÊNCIA RELATIVA DE FEMINICÍDIOS POR DIA DA SEMANA AGRUPADOS POR ANO



FONTE: O Autor (2022).

O atributo “diasemana”, cadastrado nessa base tem um padrão que pode ser observado nos anos de 2018 e novamente em 2019, as mortes por feminicídio ocorrerem principalmente no domingo, ou “Dom”, como representado no GRÁFICO 7.

No ano de 2018 o domingo se destaca, mas quinta-feira ficou com 16% em segundo lugar em incidência, fato que não se repetiu nos outros anos. O final de semana (sexta, sábado e domingo) representa 57% das mortes em 2018, 58% em 2019 e 51% em 2020.

No ano de 2019 o final de semana se destaca dos outros dias da semana que tem proporções mais parecidas, com diferença de até 3%.

No ano de 2020 segunda-feira teve um comportamento diferente dos outros anos ficando em terceiro lugar na ordem, representando 19%, fato que pode ser considerado inesperado no comportamento dos dados como a quinta-feira em 2018.

No intuito de observar a relevância dos feriados nas mortes por dia da semana, excluiu-se a semana que contava com feriados nacionais dos três anos. Como exemplo, na semana de 7 de setembro de 2019 foram excluídos dos dados todas as mortes ocorridas entre 2 e 8 de setembro de 2019. A exclusão foi realizada por script no RStudio.

Verificou-se que aproximadamente 81% das mortes ocorrem em semanas sem feriados em 2018, em 2019, 69% e 2020, 71%, sendo a média aproximada de 74% nos três anos.

Nos sábados (3%) e nas terças (2%) ocorrem mais mortes em semanas sem feriados em 2018, por outra via, domingo (3%), quarta (1%) e quinta (2%) ocorrem menos mortes para o mesmo ano. Portanto, domingo, quarta e quinta sofrem influência dos feriados nas mortes de 2018.

Nos domingos (1%), nas segundas (2%), nas quartas (1%) e quintas (2%) ocorrem mais mortes em semanas sem feriados para o ano de 2019. No entanto, nas terças (7%) ocorrem menos mortes para o mesmo ano. Então, ocorrem mais mortes em semanas com feriados nas terças de 2019, maior diferença verificada na comparação.

Nas sextas (3%) e nos sábados (3%) ocorrem mais mortes em semanas sem feriados, porém, nas segundas (5%) e quintas (1%) ocorrem menos mortes para o ano de 2020. As segundas e quintas em 2020 sofrem mais influência dos feriados nas mortes.

Já o atributo que representa o grau de instrução das vítimas, apresentando completude abaixo de 10%, teve a apresentação gráfica descartada.

Através da estatística descritiva apresentada para a base SCOL, observa-se que a vítima de feminicídio no Paraná: tem em média 37 anos completos, sendo que 25% estão abaixo de 27 anos e 25% estão acima de 43 anos, com a ressalva da dispersão elevada; tem o seu autor conhecido (ou identificado); falece principalmente vítima de arma branca (40%), arma de fogo (27%) ou agressão física (9%); é convivente (36%) ou casada (12%); trabalha com serviços domésticos (fora ou dentro de casa) (28%), está desempregada (22%) ou serviços gerais (9%); tem cor da pele branca (57%) ou parda (28%) e; sofre a violência fatal no fim de semana (sexta, sábado e domingo) com destaque para o domingo (25%) e sábado (19%). Foi também verificado a existência de vítimas idosas, 7 vítimas entre 61 e 69 anos, 4 vítimas entre 72 e 79 anos e uma vítima de 83 anos. As vítimas acima de 60 anos representam 4,6% do total de vítimas.

Na revisão de literatura sobre análise de dados e violência doméstica não foram encontrados resultados similares aos encontrados nessa base. Muitas bases usadas naquelas pesquisas não continham atributos parecidos ou não estudavam vítimas fatais.

4.1.2 Estatística descritiva da base BOU

Na base BOU foram observados valores acima de 100 anos em 43 linhas no atributo “idadeestimada”, fato que pode ser interpretado como erro de digitação, incluindo valores como 191, 343 e 1819. Os valores acima de 99 anos foram excluídos na realização da estatística descritiva. Os cálculos foram realizados via script no RStudio, incluindo a exclusão dos valores citados acima.

Como já foi citado anteriormente, uma ocorrência pode ser representada na base com mais de uma linha se ocorrer o cadastro de mais de uma natureza, meio empregado ou grupo meio empregado. As linhas dessa base contêm vítimas que registraram boletim em mais de uma ocasião, portanto pode se observar na base um mesmo indivíduo que registrou boletins em anos diferentes (2018, 2019, 2020), obviamente com idade diferente, podendo variar no máximo em 4 anos do primeiro registro para o último. A estatística descritiva deste atributo é vista na TABELA 3.

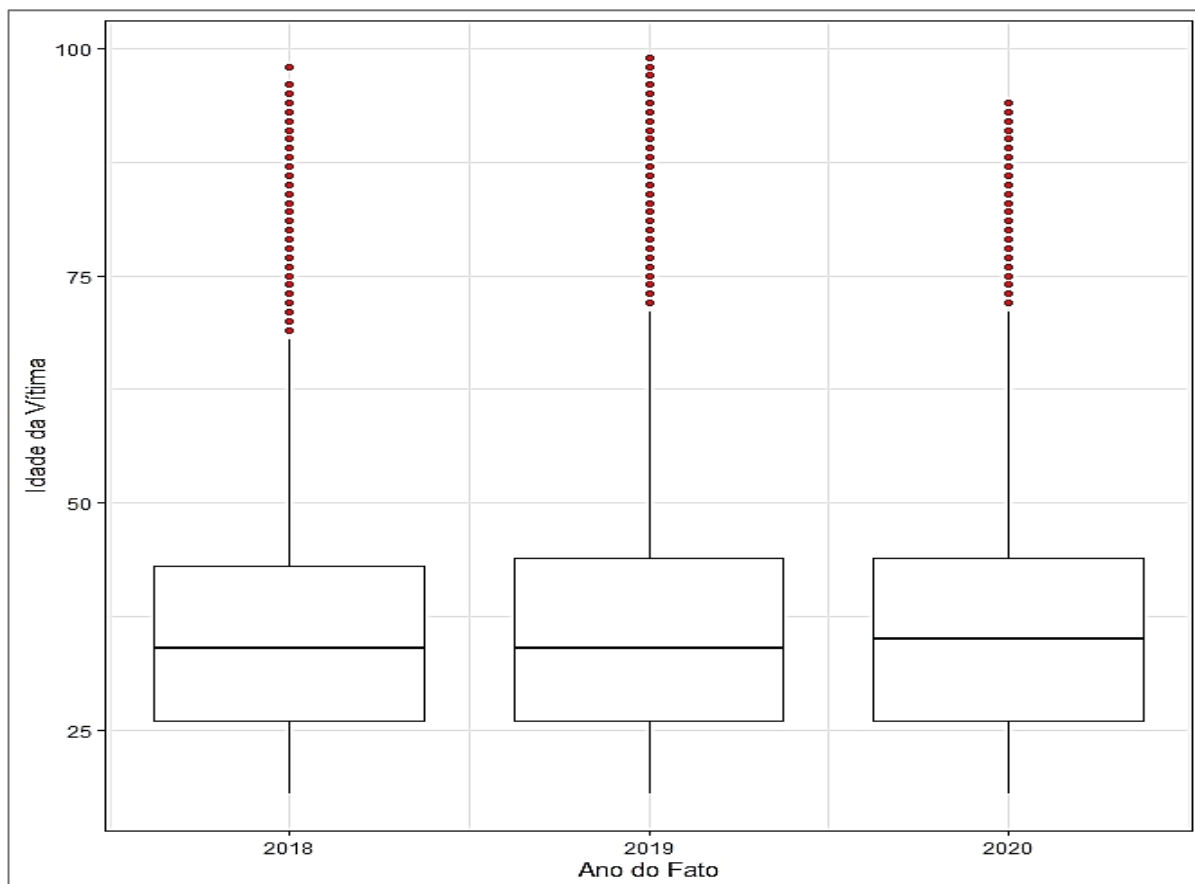
TABELA 3 – ESTATÍSTICA DESCRITIVA DAS IDADES ESTIMADAS REGISTRADA NO BOU AGRUPADAS POR ANO

Ano	Média	Desvio Padrão	Coefficiente de Variação (%)	1º Quantil	3º Quantil
2018	35,91	12,81	35,66	26	43
2019	36,34	12,96	35,67	26	44
2020	36,59	13,03	35,62	26	44

FONTE: O Autor (2022).

A média das idades é bem próxima a 36 anos. O desvio padrão aproximado é de 13 anos. Das vítimas de VD do BOU, 25% têm até 26 anos e outros 25% tem mais de 43 anos. Os coeficientes de variação observados nessa distribuição sugerem a heterogeneidade nessa distribuição.

GRÁFICO 8 - COMPARATIVO DE IDADES ESTIMADAS DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA NO BOU AGRUPADAS POR ANO

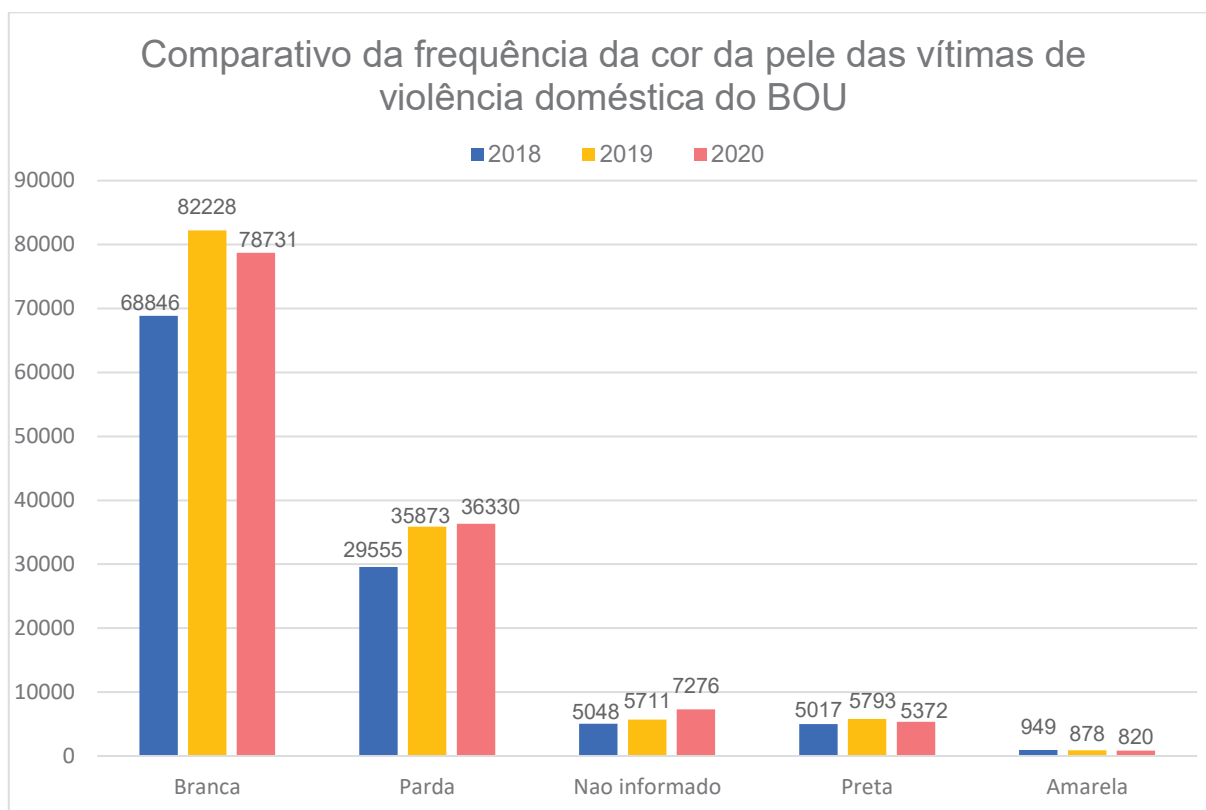


FONTE: O Autor (2022).

O valor mínimo dessa distribuição de idades é de 18 anos e o máximo é de 99 anos. Os *outliers* do GRÁFICO 8 são representados pelos pontos vermelhos. Nota-se que as vítimas de violência doméstica com idade maior de 65 anos existem em número suficiente para criar uma linha que cobre quase toda extensão superior do gráfico nos três anos. Das 368.432 linhas com idade de 18 a 99 anos, 13.317 tem o valor entre 65 e 99. Portanto, as vítimas maiores de 65 anos representam 3,61% dessa distribuição.

A base BOU contém um atributo para representar a cor da pele das vítimas. Foi observado que a ordem de proporções é igual: “Branca” > “Parda” > “Preta” > “Amarela”. Foram excluídas 49 linhas onde o valor era nulo para criação do GRÁFICO 9, sendo 7 delas em 2019 e 42 delas em 2020.

GRÁFICO 9 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DA COR DA PELE DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADAS POR ANO



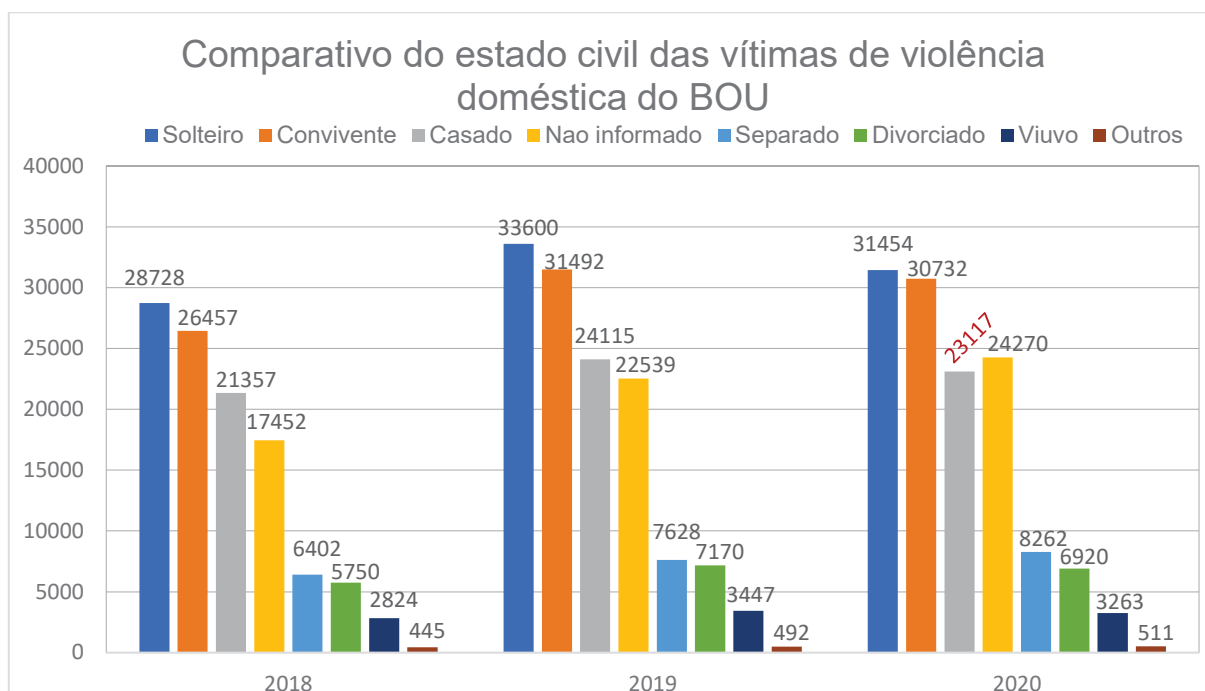
FONTE: O Autor (2022).

Observa-se o aumento em todos os anos dos valores “Não informado” e “Parda”, mesmo no ano de 2020 quando o valor total de ocorrências diminuiu 1,5% comparado ao ano anterior. A cor da pele “Branca” sofre variação acompanhando o total de ocorrências dos anos. “Amarela” teve uma queda ordenada de 7% durante o período e “Preta” teve crescimento de 2018 para 2019, porém teve queda em 2020, mesmo período que “Não informado” teve o crescimento de 27%.

O “Não informado” representou aproximadamente 5% da distribuição.

Um atributo que representa o estado civil da vítima, “estadocivil”, na base BOU continha 49 linhas com valor nulo. Essas linhas foram excluídas na criação do GRÁFICO 10.

GRÁFICO 10 – COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO ESTADO CIVIL DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADOS POR ANO



FONTE: O Autor (2022).

O total de linhas preenchidas para esse atributo foram 109.415 em 2018, 130.483 em 2019, 128.529 em 2020.

A proporção de registros de estado civil segue a ordem representada na legenda do gráfico com a exceção de “Casado” em 2020, valor representado em vermelho no GRÁFICO 10 (23.117). Nesse caso os registros de “Não informado” superaram os registros de “Casado”.

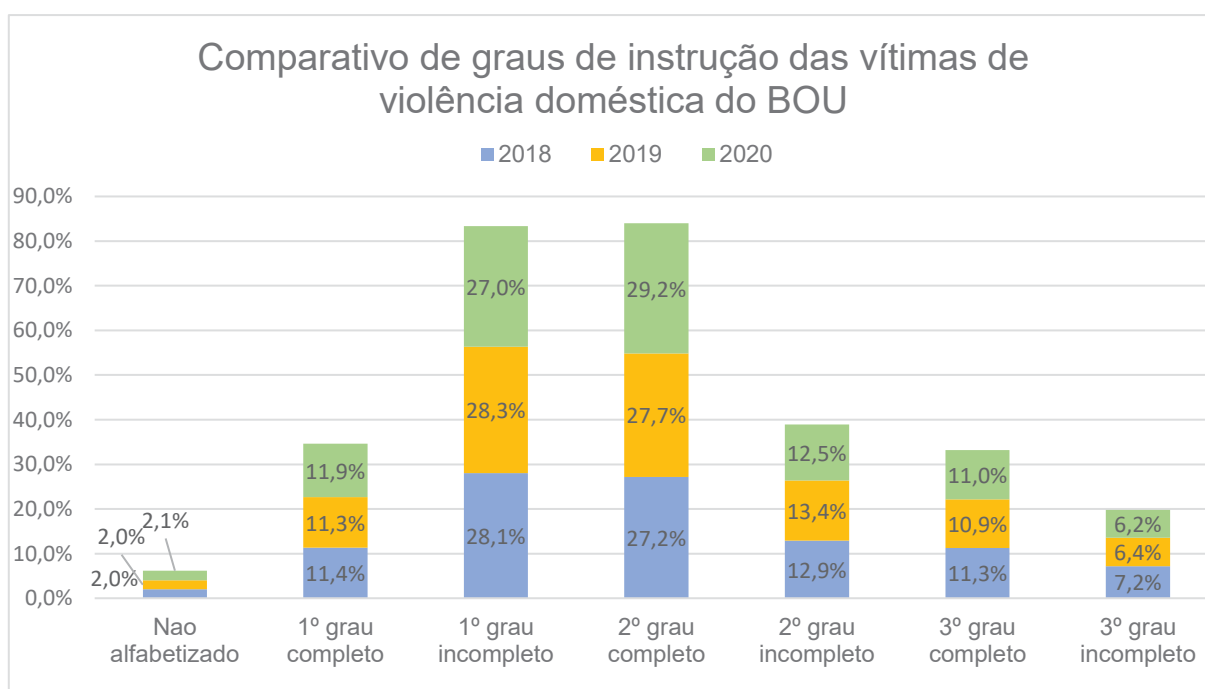
Os valores “Solteiro”, “Convivente”, “Casado”, “Divorciado” e “Viuvo” tem um comportamento que acompanha o valor total de ocorrências para cada ano, cresce em 2019 e tem queda em 2020, porém, “Solteiro”, “Casado” e “Viuvo” apresentaram uma queda percentual entre 4% e 6%, ou seja, queda maior que a variação negativa do total de 1,5%.

No caso dos valores “Não informado”, “Separado” e “Outros” houve crescimento independentemente da variação do valor total de ocorrências no período, crescendo 7,68%, 8,31% e 3,86% respectivamente no ano de 2020, onde os outros valores apresentaram queda.

Os três primeiros valores, “Solteiro”, “Convivente” e “Casado” representam 68% de toda a contagem no período.

Nessa base foi verificado que o atributo “grau_instrucao” que representa o grau de instrução das vítimas do BOU continha poucos campos vazios, o que tornou a comparação gráfica possível, conforme apresentado no GRÁFICO 11.

GRÁFICO 11 – COMPARATIVO DA FREQUÊNCIA RELATIVA DO GRAU DE INSTRUÇÃO DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADOS POR ANO



FONTE: O Autor (2022).

Em 2018, essa base continha 109.415 registros válidos para esse atributo, 130.470 em 2019 e 128.519 em 2020, o que representou 99,9% de todas as ocorrências. No entanto, foram excluídas da representação gráfica 31% (média dos três anos) das linhas com o valor “Não informado”, visto que, tal valor não auxilia na generalização desse atributo.

Os valores “Não alfabetizado”, “1º grau completo”, “2º grau completo” e “3º grau completo” apresentaram crescimento não acompanhando a variação do total de ocorrências. O valor “3º grau incompleto” sofreu queda gradual nos três anos. Os valores “1º grau incompleto” e “2º grau incompleto” tiveram um comportamento de crescimento em 2019 e queda em 2020. De 2018 para 2019 existe um crescimento de 21.075 no total de registros (+19%) e de 2019 para 2020 houve uma queda de 3.800 registros (-2,91%). O comportamento de crescer em 2019 e decrescer em 2020 seria o esperado caso não houvesse nenhum valor se destacando no período.

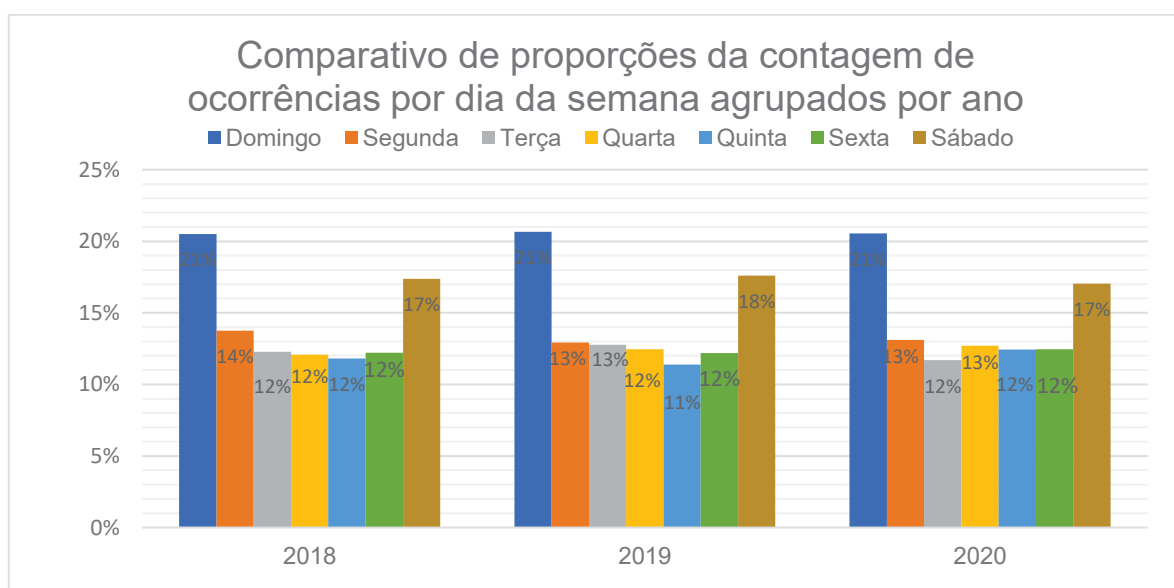
Um fato observado na frequência absoluta desse atributo foi a queda de boletins em 2020 para os valores com sufixo “incompleto”. Uma possível explicação seria a conclusão dos cursos pelas vítimas durante o período de três anos, portanto, são as mesmas vítimas. Outra possível explicação se apoiaria na mudança de perfil das vítimas para esse atributo no ano de 2020. Os valores “2º grau completo” e “1º grau incompleto” representam aproximadamente 56% da distribuição, sendo os mais frequentes.

Da mesma forma que foi visto na base de dados analisada anteriormente, essa base continha o atributo “diasemana”, calculado através do atributo “fatoinical”, esse representa o dia da semana da ocorrência de violência doméstica no BOU.

O atributo não continha linhas vazias ou com conteúdo nulo, tendo 109.415 registros em 2018, 130.490 em 2019 e 128.571 em 2020.

A distribuição gráfica pode ser verificada no GRÁFICO 12 abaixo.

GRÁFICO 12 - COMPARATIVO DE FREQUÊNCIA RELATIVA DE BOLETINS DO BOU POR DIA DA SEMANA AGRUPADOS POR ANO



FONTE: O Autor (2022).

É possível observar nesse comparativo que o “Domingo” e o “Sábado” são as maiores proporções dessa distribuição. Se somados representam 38% da contagem em 2018 e 2020, ainda, 39% em 2019. Esses dois dias da semana foram bem constantes com relação a proporção nesse período, fato também visto nos dias do meio da semana. Com exceção de domingo e sábado, a média dos dias da semana

foi aproximadamente de 12% no período. Portanto, o comportamento das ocorrências de violência doméstica por dia da semana foi quase idêntico nos anos representados nesse gráfico.

Na base SCOL esse atributo se comportou um pouco diferente nos dias da semana “Quinta” em 2018 com a proporção de 16% e “Segunda” em 2020 com 19% e “Sexta” em 2019 com 19%, proporções acima em pelo menos 3% das vistas na base BOU para o meio da semana. Em 2020 também se destacaram por apresentarem proporções muito baixas em comparação aos 12% do BOU para o meio da semana, “Quarta” e “Sexta” apresentaram 5% e 7% respectivamente. Ainda, em 2018 também apresentaram proporção de 7% os dias “Segunda” e “Terça”.

Comparando as representações gráficas desse atributo nas duas bases observa-se um padrão muito mais claro na base BOU para os dias “Sábado” e “Domingo”, o meio da semana no período varia menos nessa base.

No intuito de observar a relevância dos feriados nas ocorrências por dia da semana, excluiu-se a semana que contava com feriados nacionais dos três anos no RStudio. Verificou-se que em média 77% das ocorrências são registradas em semanas sem feriados.

Observou-se proporções similares para as semanas com e sem feriados. A sexta-feira (1%) em 2018 e a terça-feira (1%) em 2019 foram influenciadas pelas semanas com feriados, em outra via, os sábados e domingos (1%) tem maior registro nas semanas sem feriados em 2020. A ordem das proporções não mudou nas semanas com e sem feriados.

Essa base ainda conta com um atributo para ocupação das vítimas, “ocupacao”, que tem o conteúdo idêntico do “profissao” da base SCOL. A quantidade de profissões diferentes a cada ano nessa base dificulta a representação gráfica de todas. Em 2018 foram 260 valores diferentes, em 2019 foram 266 e 2020 foram 252.

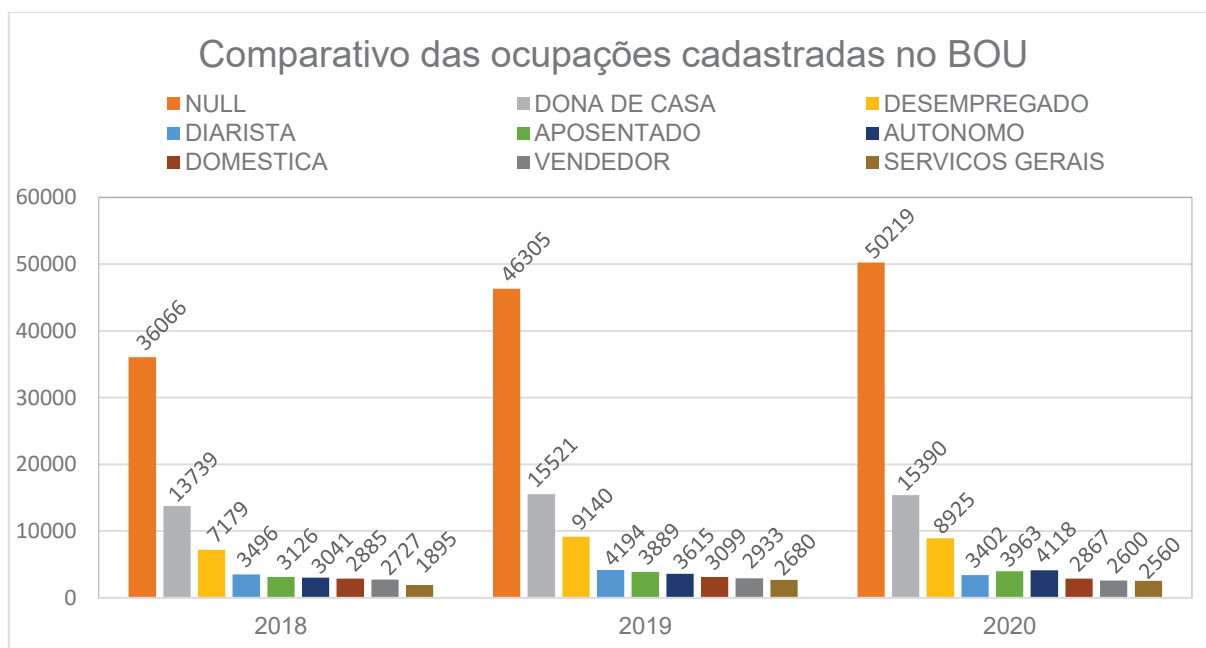
Na observação das proporções que representavam na contagem agrupada por ano, verificou-se que as nove primeiras ocupações ali registradas se repetiram nas primeiras posições todos os anos e representou 68% da distribuição em 2018, 70% em 2019 e 73% em 2020.

As ocupações “ESTUDANTE” e “ATENDENTE” apareceram entre os onze primeiros em um dos três anos, “AUXILIAR DE PRODUÇÃO” ficou entre os dez primeiros em 2018 e entre os onze primeiros em 2020. Portanto, a inconstância

dessas três nas primeiras posições orientou a escolha para representar somente as primeiras nove.

A contagem agrupada por ano foi representada no GRÁFICO 13 abaixo.

GRÁFICO 13 - COMPARATIVO DA FREQUÊNCIA ABSOLUTA DA OCUPAÇÃO DAS VÍTIMAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADAS POR ANO



FONTE: O Autor (2022).

O primeiro fato a ser notado nesse gráfico é que o valor nulo “NULL” é o maior em todos os anos. Outro fato é que esse valor cresceu em todos os anos independente do comportamento de crescimento de 19% em 2019 e queda de 1,5% em 2020, baseando-se nos valores totais para cada ano: 109.415, 130.490 e 128.571.

Outros valores para o atributo que cresceram independentemente do comportamento dos totais de cada ano foram: “APOSENTADO” e “AUTONOMO”. Como todas as ocorrências selecionadas na extração dessa base são criminais, nota-se um crescimento da violência contra o idoso. Quanto ao crescimento dos autônomos na contagem, o fato pode estar relacionado à pandemia do “covid19” no fim de 2019 e todo ano de 2020 onde muitas empresas fecharam aumentando o desemprego no estado. Esses dois fatos poderiam ser confirmados em um estudo específico para cada observação dessa.

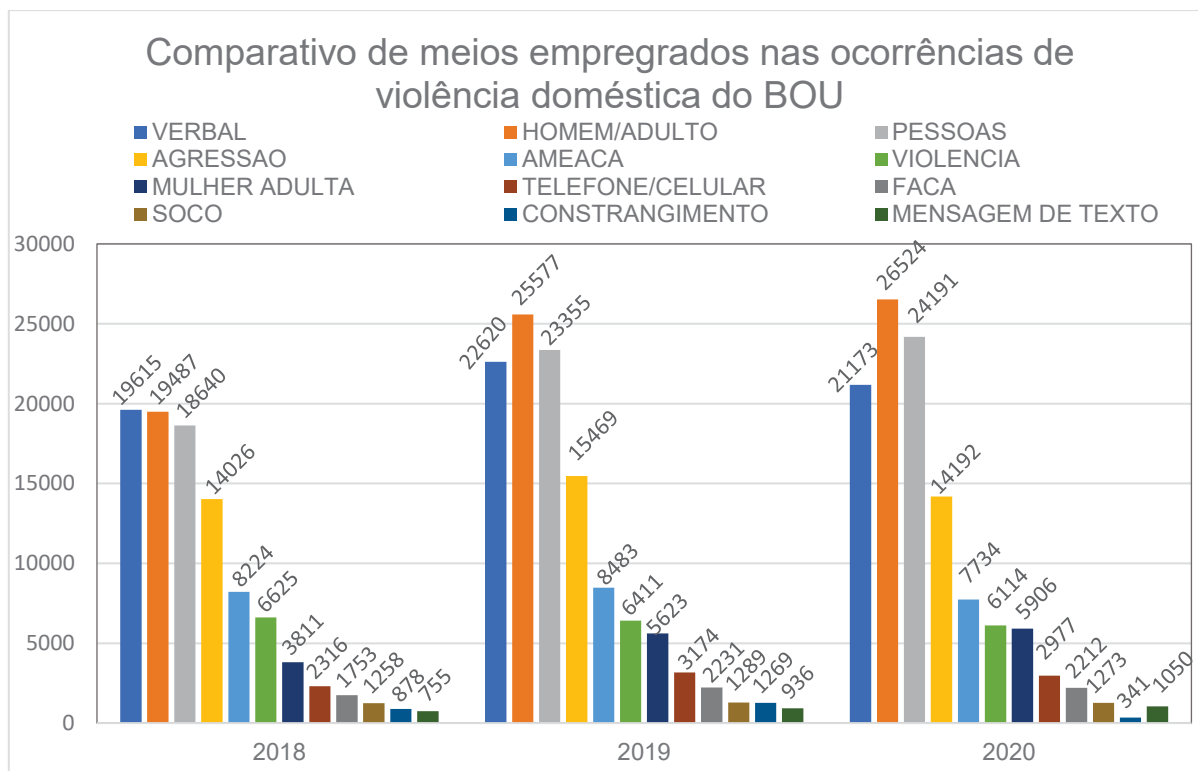
Acompanhando o comportamento de crescimento em 2019 e queda em 2020 estão os valores: “DONA DE CASA”, “DESEMPREGADO”, “DIARISTA”, “DOMESTICA”, “VENDEDOR” e “SERVICOS GERAIS”.

Nota-se também que o valor “DIARISTA” é o que teve variação proporcional de crescimento e queda mais agudos que os outros valores.

Na base SCOL é possível registrar um meio empregado (principal meio empregado que causou a morte), nessa base, BOU, o atributo “meio_empregado” armazena um ou mais valores para cada boletim. Notou-se no agrupamento dos meios empregados por ano que, excluindo o valor “NÃO DEFINIDO” os doze valores mais frequentes em todos os anos se repetiam. Esses doze primeiros valores em 2018 representaram aproximadamente 90% em 2020 e 91% da contagem em 2018 e 2019.

O comportamento de crescimento do valor total de 2018 para 2019 foi de 19% e de queda em de 2019 para 2020 foi de 1,3%, valores parecidos com alguns atributos já representados graficamente neste.

GRÁFICO 14 - COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO MEIO EMPREGADO NAS OCORRÊNCIAS DE VIOLÊNCIA DOMÉSTICA DO BOU AGRUPADOS POR ANO



FONTE: O Autor (2022).

Observou-se no GRÁFICO 14 o crescimento independente do comportamento do total de registros desse atributo nos valores: "HOMEM/ADULTO", "MULHER ADULTA", "PESSOAS" e "MENSAGENS DE TEXTO". O valor "VERBAL" foi o de maior contagem em 2019, porém "HOMEM/ADULTO" cresceu mais nos anos e representou aproximadamente 20% da contagem geral contra 17% do verbal.

Para cada boletim é obrigatório o cadastro de pelo menos um meio empregado, porém é comum que em ocorrências de violência doméstica existam pelo menos dois valores cadastrados. O valor: "MULHER ADULTA" era esperado já que a base é formada por vítimas mulheres maiores de idade. Costumam ser cadastrados em conjunto com outros meios os valores: "HOMEM/ADULTO" e "PESSOAS", identificando que pelo menos um homem adulto participa da ocorrência, como a base é de vítimas mulheres esse homem pode ser autor ou testemunha, sendo mais provável que seja autor.

Apesar do comportamento do valor "AGRESSAO" acompanhar o valor total de ocorrências para cada ano, notou-se que no geral esse meio empregado ocorre aproximadamente em 7% dos casos. Mesmo que a proporção represente menos que os valores como o "VERBAL", alguns feminicídios da base SCOL tem agressão como o meio empregado para a morte. Isso vale também para os valores "FACA" e "SOCO", já que o principal meio empregado nos feminicídios é arma branca ou imprópria. A faca é o exemplo mais prático para explicar o que pode ser uma arma branca.

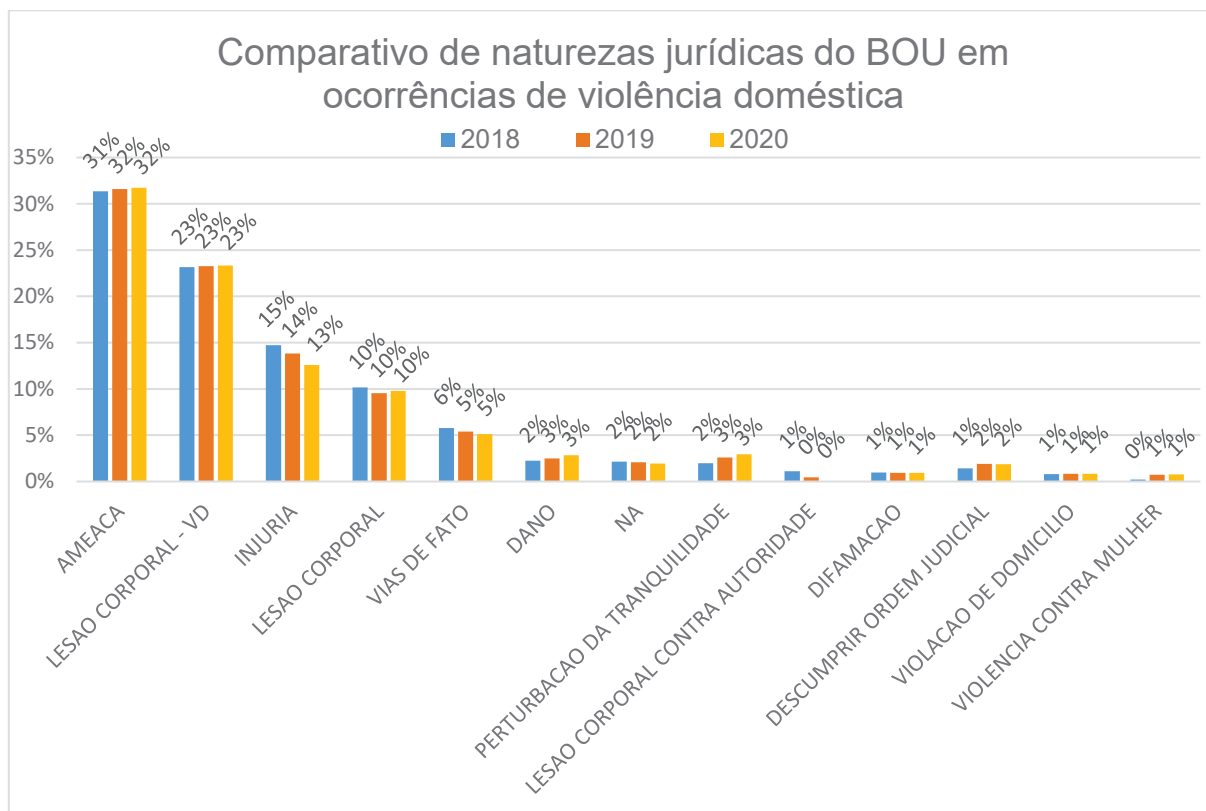
O crescimento do valor: "MENSAGENS DE TEXTO" pode ser considerado uma novidade nessa contagem. À medida que as vítimas são orientadas pelos órgãos de proteção ao cidadão (polícias estaduais, guarda municipal, ministério público) ou até orientadas por outras vítimas, ficam avisadas que fatos de violência doméstica por meio eletrônico ficam registrados para que possam ser transcritos nos boletins de ocorrência. Portanto, esse crescimento anual pode ter ocorrido por esses motivos.

No caminho inverso às mensagens de texto o valor "VIOLENCIA" decresceu, um fato positivo para a vítimas de violência doméstica no estado.

O valor "CONSTRANGIMENTO" teve comportamento de crescimento e depois queda, porém a queda em 2020 foi proporcionalmente a mais aguda. Observação que também pode ser comemorada pelas vítimas.

Os demais valores tiveram crescimento no ano de 2019 e queda no ano 2020 acompanhando dentro da razoabilidade o comportamento dos valores totais para cada ano.

GRÁFICO 15 - COMPARATIVO DA FREQUÊNCIA ABSOLUTA DO NOME JURÍDICO EM OCORRÊNCIAS DE VIOLÊNCIA DOMÉSTICA AGRUPADOS POR ANO



FONTE: O Autor (2022).

As naturezas jurídicas representadas nesse gráfico estiveram entre os primeiros lugares na contagem de todos os anos. O atributo “nomejuridico” teve 140 valores diferentes em 2018, 144 em 2019 e 156 em 2020, o que dificulta a representação de todos graficamente, como o caso do atributo “meio_empregado”.

Portanto as 12 naturezas (mais NA) que compõem o GRÁFICO 15 representaram aproximadamente 96% da contagem em 2018 e 2019 e 95% em 2020. Importante ressaltar que os valores descritos no gráfico como “NA” representam dados não disponíveis. Os dados foram importados e transformados na ferramenta RStudio e quando o atributo tem valor vazio o programa substitui por “NA” ou “Not Available” do inglês.

O valor “DESCUMPRIR MEDIDA JUDICIAL” foi criado através da soma de “DESOBEDIENCIA A DECISAO JUDICIAL”, “DESCUMPRIR DECISAO JUDICIAL QUE DEFERE MEDIDAS PORTETIVAS DE URGENCIA” e “DESCUMPRIR ORDEM JUDICIAL DE MEDIDA PROTETIVA” pois as naturezas jurídicas servem para o mesmo propósito.

Nota-se que ameaça, lesão corporal – violência doméstica e lesão corporal apresentam valores bem constantes durante os anos. Ameaça com aproximadamente 32% e as lesões corporais somadas com 33%. Também foi o comportamento de algumas naturezas de menor proporção como o caso de difamação e violação de domicílio com 1% em todos os anos.

Injúria teve uma queda ordenada no período, mas representa 14% em média nos anos, percentual que, somado com as médias de ameaça e as lesões representa 79% da contagem agrupada do período.

O número de “NA” foi constante em 2%, o melhor dos cenários para análise de dados é que todos os dados estejam disponíveis, porém, um valor constante dessa proporção tende a não interferir negativamente nos resultados.

Vias de fato apresentou leve queda ficando na casa dos 5%. Essa natureza é registrada quando existe uma luta corporal entre duas pessoas sem que seja possível identificar quem é o autor ou a vítima, pois os dois envolvidos agredem.

A lesão corporal contra autoridade ou agente quase zerou de 2018 para 2019 e 2020. Esse valor é cadastrado quando alguma autoridade ou agente do governo que vai prestar atendimento a alguma vítima é agredido no curso do atendimento.

Verificou-se leve crescimento nas naturezas de perturbação da tranquilidade, dano, descumprir medida judicial e violência contra mulher, todos na casa de um 1%. A “violência contra mulher” não pode ser somada com ameaça, lesão corporal, injúria ou qualquer outro tipo de natureza, já que a natureza é parte da lei maria da penha, o que engloba todas as ações descritas nela como violência doméstica e familiar.

O aumento da natureza de violência contra mulher pode ter ocorrido devido a algumas ocorrências necessitarem de mais de um tipo de natureza jurídica para descrever os fatos e esses estarem descritos na lei maria da penha. Se uma ocorrência a mulher impedida de sair de casa, foi xingada e teve alguma roupa destruída pelo cônjuge, o policial cadastraria a natureza “violência contra a mulher”.

Antes de existir essa lei, Maria da Penha, o cadastro das naturezas nesse boletim de ocorrência poderia ter valores como: dano, injúria e cárcere privado. Essas naturezas em conjunto descreveriam a conduta mais detalhadamente, porém, dependendo da forma da recuperação dos dados e o contexto da apresentação dos dados pode-se somar o dano com outros dessa natureza e agrupá-los como crimes contra o patrimônio. A existência da lei Maria da Penha no sistema BOU auxilia o agrupamento de dados quando tratamos de violência doméstica.

Quanto ao crescimento do descumprimento de decisão judicial, aproximadamente 1% de 2018 para 2019 e 2020, uma explicação possível é o crescimento de vítimas que procuraram as delegacias da mulher ou o ministério público afim de solicitar medida protetiva de urgência contra um ou mais autores de violência doméstica. Essa natureza será cadastrada se a vítima tem a medida protetiva solicitada e um autor descumpriu o previsto na medida protetiva. Então, para que se descumpra a medida protetiva é necessário já tê-la solicitado, ou seja, o cadastro dessa natureza depende da emissão desse documento anteriormente pelos órgãos competentes.

O caso da perturbação da tranquilidade e do dano, na maioria das vezes não está relacionado à violência doméstica. Porém, as vítimas podem ser perturbadas pessoalmente no trabalho, por meio eletrônico (telefone, mensagem de texto, redes sociais) ou em sua residência. Também, vítimas quando contrariam autores frequentes de violência doméstica em algumas ocorrências tem bens destruídos como punição ou durante uma discussão mais violenta. Por isso essas duas naturezas representam no máximo 3% no período.

O padrão de ocorrências de violência doméstica para “nomejuridico” verificado nesse período pode ser sintetizado pelas naturezas jurídicas: ameaça, lesão corporal, lesão corporal – violência doméstica e injúria. Duas delas envolvem violência psicológica e as outras duas, violência física. Não foi possível comparar os resultados da base BOU com a SCOL pois a base de dados de ocorrências letais usada anteriormente é formada exclusivamente por vítimas da natureza “feminicídio”.

Na observação da estatística descritiva apresentada sobre os atributos dessa base pode-se descrever a vítima de violência doméstica mulher do Paraná registrada no BOU como: adulta de idade média de 36 anos, tendo 25% dessas abaixo de 26 anos e 25% com pelos menos 43 anos, com a mesma ressalva da alta dispersão encontrada na base SCOL; tem a cor da pele branca (62%) ou parda (28%); tem como principais estados civis solteiro (25%) ou convivente (24%), seguido por casado (19%); o grau de instrução mais provável é 1º grau incompleto (28%) ou 2º grau completo (28%); tem a sua ocupação ligada ao serviço doméstico em casa ou fora de casa (15%), pode estar desempregada (7%) ou é aposentada (3%); sofre a violência doméstica principalmente aos finais de semana (sábado e domingo); sofre violência verbal (18%), agressão (12%) e ameaça (7%) como principais meios empregados e

sofre violência física em aproximadamente 21% dos casos de maior contagem, os que compõem o GRÁFICO 15.

Observando as principais naturezas jurídicas, o que descreve a(s) conduta(s) criminosa(s) de cada boletim de ocorrência dessa base, observa-se que em aproximadamente 39% das vítimas de VD no Paraná sofrem violência física e aproximadamente 47% sofrem das violências verbal ou psicológica.

4.1.3 Estatística descritiva da base SCOL *versus* BOU

A análise estatística de alguns atributos das bases SCOL e BOU ofereceu uma visão do comportamento dos dados de cada uma delas. Na base SCOL, 6 atributos foram representados graficamente, na base BOU, 8 desses. Não seria possível comparar todos os atributos das bases, porém, os seis atributos representados graficamente da base SCOL também existem na base BOU. O QUADRO 17, oferece uma comparação da descoberta da estatística descritiva nas duas bases para os atributos coincidentes.

Nota-se que quanto as idades a base SCOL tem média de 37 anos e BOU 36 anos. Tem-se 25% com 26 anos ou menos no BOU, quando que na outra base 28 anos, o que é compatível com as suas respectivas médias. Apesar da base SCOL apresentar números ligeiramente maiores para média de idade, 25% das vítimas dessa tem mais de 44 anos, quando que na base BOU tem mais de 43 anos. De qualquer forma, no quesito idade das vítimas os valores calculados são parecidos, o que não facilita a encontrar padrões muito distintos nas bases.

O número de vítimas com idade acima de 65 anos nas bases também foi proporcionalmente bem parecido, 3,61% e 3,81%. A maior diferença para o atributo idade nas bases foi a Moda, tendo 29 anos para o SCOL e 24 anos para o BOU. No QUADRO 16, para a moda do atributo idade no BOU foi adicionado 23 anos juntamente com 24 anos, pois a diferença entre eles foi de 0,03% do total, portanto, o valor é quase idêntico para as duas idades. As idades seguintes, do maior para o menor foram 26 e 25 anos, demonstrando que as vítimas de VD do BOU são mais jovens comparadas as vítimas fatais do BOU.

QUADRO 16 - COMPARATIVO DA ANÁLISE ESTATÍSTICA DAS BASES BOU E SCOL

ATRIBUTOS DAS BASES	BOU	SCOL
Idade estimada	36 anos	37 anos
Idade 25% abaixo dos	26 anos	28 anos
Idade 25% acima dos	44 anos	43 anos
Idades acima de 65 anos	3,61%	3,81%
Moda da Idade	23/24 anos	29 anos
Meios Empregados (ordem)	Verbal > Agressão > Ameaça	Arma Branca ou Imprópria > Arma de Fogo > Agressão Física
Cor da pele (ordem)	Branca > Parda > Preta > Amarela	Branca > Parda > Preta > Amarela
Profissão/Ocupação (ordem)	Serviços domésticos > Desempregada > Aposentada > Autônoma > Vendedora > Serviços gerais	Serviços domésticos > Desempregada > Serviços Gerais > Estudante > Agricultora > Professora
Estado civil (ordem)	Solteira > Convivente > Casada > Separada > Divorciada > Outros	Convivente > Casada > Solteira > Separada > Divorciada > Outros
Dia da semana (ordem)	DOM > SAB > SEG > QUA > SEX > TER > QUI	DOM > SAB > SEG > SEX > QUI > TER > QUA

FONTE: O Autor (2022).

Os meios empregados das bases diferem muito pois as vítimas do SCOL são fatais, então seus meios empregados serão muito mais violentos ou letais. Porém, o segundo item do BOU, Agressão, também aparece no SCOL. Infelizmente observa-se que agressão é um fato comum para as vítimas de ambas as bases.

A cor da pele informada para as vítimas seguiu a mesma ordem de proporção na comparação.

Na profissão ou ocupação, serviços domésticos (resultados da soma de dona de casa, doméstica, diarista nas bases) e desempregada são as mais frequentes nas duas bases. Também, Serviços gerais é uma ocupação que coincide nessas. Como na estatística existem muito poucas coincidências, principalmente em caso de análises de frequência relativa, três ocupações representam bem as vítimas das duas bases.

No atributo estado civil observa-se uma diferença na ordem das frequências. As vítimas do BOU em sua maioria Solteira ou conviventes, quando que no SCOL as

conviventes e casadas são as vítimas mais frequentes. As vítimas conviventes aparecem entre os dois valores mais frequentes nas duas bases, porém, nas duas bases é o valor mais propenso ao feminicídio.

A frequência relativa do atributo dia da semana teve ordem de proporções idênticas nos três primeiros lugares. Tanto os feminicídios quanto as ocorrências de VD ocorrem mais concentradas no Domingo, Sábado e Segunda-feira. A exclusão das semanas com feriados nacionais não alterou a proporção dos primeiros lugares demonstrados no QUADRO 16, porém, para os dados do BOU a ordem do quinto e sexto lugares (sexta e terça) foi invertida, assim como o quarto e quinto lugares do SCOL (sexta e quinta) sem feriados nacionais.

Alguns atributos das ocorrências de VD e feminicídio foram bem parecidos, tendo a moda das idades como a característica que mais as diferenciou. Conviventes são mais suscetíveis a ocorrência fatal, a agressão física é presente para os dois lados da comparação, vítimas com ocupação em serviços domésticos ou desempregada são mais frequentes nas ocorrências e nas mortes, a maioria das mortes e das ocorrências de VD ocorrem no fim de semana e na segunda-feira.

4.1.4 Estatística descritiva da Base Naturezas *versus* feminicídios

A Base Naturezas foi solicitada para que fosse possível observar o comportamento dos dados das vítimas de violência doméstica em relação as naturezas jurídicas dos boletins de ocorrência. As naturezas mais graves que podem ser vistas nessa base estão relacionadas à morte das vítimas, como: Feminicídio, Homicídio (Simples ou Qualificado), Lesão corporal seguida de morte, Atendimento de local de morte e Morte a apurar com indícios de crime.

O tratamento e seleção dos dados das vítimas de feminicídio do SCOL (236 vítimas) retornou à contagem das naturezas jurídicas e boletins de 149 dessas, o que representa aproximadamente 63% de todas as vítimas do SCOL. As outras 87 vítimas não foram encontradas nessa base muito provavelmente devido a sua morte não ter sido relacionada à violência doméstica (ou feminicídio) quando a ocorrência foi atendida pelas polícias, então, o registro do SCOL foi modificado após o início das investigações. Portanto, todas as naturezas jurídicas registradas nos anos de 2018, 2019 e 2020 com a política pública “VIOLENCIA DOMESTICA” das 149 vítimas foi computada.

Foram registradas 26 naturezas jurídicas diferentes nos 239 boletins de ocorrências das 149 vítimas selecionadas. As frequências absoluta e relativa das naturezas jurídicas podem ser visualizadas na QUADRO 17.

QUADRO 17 - FREQUÊNCIA DAS NATUREZAS JURÍDICAS NAS OCORRÊNCIAS DO BOU AGRUPADAS POR VÍTIMA DE FEMINICÍDIO

NATUREZAS JURÍDICAS (NOME JURÍDICO) DAS VÍTIMAS DE FEMINICÍDIO NO BOU	Freq. Absoluta	Freq. Relativa	Grupo
FEMINICÍDIO	60	20%	Mortes
HOMICÍDIO QUALIFICADO	28	10%	
ATENDIMENTO DE LOCAL DE MORTE	28	10%	
MORTE A APURAR COM INDÍCIOS DE CRIME	5	2%	
HOMICÍDIO SIMPLES	4	1%	
LESAO CORPORAL SEGUIDA DE MORTE	2	1%	
LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR	48	16%	Lesões
LESAO CORPORAL	19	6%	
LESAO CORPORAL DE NATUREZA GRAVISSIMA	9	3%	
LESAO CORPORAL DE NATUREZA GRAVE	5	2%	
LESAO CORPORAL CONTRA AUTORIDADE OU AGENTE	3	1%	
AMEACA	40	14%	Ameaça e Disparos
DISPARO DE ARMA DE FOGO	3	1%	
INJURIA	11	4%	Contra a Honra
CALUNIA	1	0%	
PERTURBACAO DA TRANQUILIDADE	5	2%	Perturbação
DESCUMPRIR DECISÃO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS	4	1%	
DESCUMPRIMENTO DE MEDIDA PROTETIVA	2	1%	
VIAS DE FATO	5	2%	Vias de Fato
CONSTRANGIMENTO ILEGAL	1	0%	
DANO	3	1%	Contra o Patrimônio
FURTO SIMPLES	2	1%	
DANO QUALIFICADO	1	0%	
ESTUPRO DE VULNERAVEL	2	1%	Crimes Sexuais
ESTUPRO OU ATENTADO VIOLENTO AO PUDOR	1	0%	
APOIO A OUTROS ORGAOS - SEM ILICITUDE	1	0%	Sem Ilícito
TOTAL DE NATUREZAS JURÍDICAS DAS VÍTIMAS DO SCOL NO BOU	293	100%	Geral

FONTE: O Autor (2022).

As 26 naturezas foram organizadas em grupos para facilitar a leitura dessa tabela. O grupo denominado “Mortes” não será usado nos processos de análise posteriores a análise de frequências, pois esse uso poderia induzir algoritmos de mineração de dados a uma resposta equivocada.

Naturezas jurídicas relacionadas ao grupo Morte, do início do QUADRO 17 até Lesão Corporal seguida de morte, foram cadastradas em 43% dos boletins, 20% foram cadastrados inicialmente como feminicídio. Para aproximadamente 37% (87 mulheres) das vítimas fatais de VD no estado do Paraná, ao terem sido cadastrados seus boletins no BOU foi selecionada uma política pública diferente de “VIOLENCIA

DOMESTICA”, um exemplo de registros que tem diferença entre o boletim inicial e a investigação.

Dessas vítimas fatais aproximadamente 29% sofreram lesão corporal antes da ocorrência de sua morte. No entanto, aproximadamente 5% sofreram lesão grave ou gravíssima, fatos que podem levar a morte da vítima posteriormente. Apesar da frequência baixa dessas lesões mais graves, essas são bem destrutivas.

Das vítimas que sofreram Lesão grave, 3 delas tiveram 1 registro e uma delas teve 2 registros da mesma natureza. Fato preocupante, pois, esse tipo de lesão pode levar a vítima a morte. Ter mais de um evento com essa capacidade destrutiva com a mesma pessoa pode ser o fato que levou a levou a falecer, já que essa vítima em questão não tem registro de nenhuma outra natureza nessa base. Uma das vítimas com um registro dessa lesão também teve 1 Injúria, 1 Lesão corporal e Morte a apurar com indícios de crime.

Observou-se que a frequência absoluta da Lesão gravíssima foi quase o dobro da Lesão grave. Foi verificado que nessa base que 8 vítimas sofreram esse tipo de lesão uma vez, sendo que uma delas sofreu desse crime 2 vezes. Essa vítima também teve registrado o feminicídio além das duas lesões. Com exceção de 3 vítimas desse crime, além do registro da Lesão gravíssima essas vítimas tiveram somente o registro de naturezas relacionadas ao grupo Morte do QUADRO 17.

A Lesão corporal foi registrada para 15 vítima uma vez e 2 vezes para duas delas. Das 17 vítimas, 3 também tinham registro de 1 Lesão corporal – violência doméstica e 2 tinham 3 registros dessa mesma natureza. Dessas vítimas, 5 registraram Ameaça, 2 Injúria, 2 Furto simples, 1 constrangimento ilegal, 1 Dano qualificado, 1 Estupro ou atentado violento ao pudor, 3 Disparos de arma de fogo, 1 Estupro de vulnerável, 1 Lesão grave e 9 naturezas relacionadas ao grupo Mortes do QUADRO 17.

Os 48 registros da natureza Lesão corporal – violência doméstica e familiar foram sofridos por 35 vítimas. Dessas, 28 sofreram o fato uma vez, 4 duas vezes, 2 três vezes e 1 seis vezes. Das 35 vítimas, 6 sofreram uma Lesão corporal de outra natureza, 2 sofreram Lesão corporal – violência doméstica e familiar em três ocasiões e uma vítima teve registro de Lesão corporal de natureza grave. Nove dessas vítimas registraram Ameaça uma vez e uma vítima registrou 3 Ameaças. Além da Lesão corporal – violência doméstica e familiar também foi verificado para essas vítimas o

registro de 3 Injúrias, 1 Descumprimento de medida judicial, 1 Perturbação da tranquilidade, 1 Furto simples e 3 Disparos de arma de fogo.

O caso de lesão contra autoridade ou agente teve aproximadamente 1% de frequência relativa nos boletins de ocorrências dessa base contendo somente vítimas fatais. A melhor explicação para essa natureza ocorrer é que, o agente ou autoridade foi lesionado no atendimento da ocorrência. Para uma das vítimas que teve o registro dessa natureza, dos três boletins associados a ela nessa base também constavam 1 Lesão corporal – violência doméstica e familiar, 1 Ameaça, 1 Perturbação da tranquilidade e 1 Descumprimento de medida judicial.

Foi observado que aproximadamente 4% das vítimas fatais registraram descumprimento de medida protetiva ou tiveram o sossego perturbado, o que pode demonstrar que uma quantidade pequena dessas vítimas solicitou medida protetiva. Outra possibilidade de interpretação é que, a baixa frequência relativa do grupo Perturbação não o torna um fator tão preponderante para a ocorrência fatal.

A natureza, Descumprir decisão judicial que defere medidas protetivas, teve 4 registros, um registro para cada vítima. Três dessas vítimas registram Ameaça, 1 uma vez, outra duas vezes e uma dessas ainda registrou 3 ameaças. Das duas vítimas que tiveram registro da natureza Descumprimento de medida protetiva, 1 teve também Ameaça e Dano e a outra teve Perturbação da tranquilidade.

Ameaça e disparo de arma de fogo juntos representaram 15% da frequência relativa. Os disparos representaram aproximadamente 1%, portanto a ameaça foi mais frequente nos casos anteriores à morte da vítima, ou seja, foi mais representativa que os disparos. Dessas vítimas, 35 tiveram pelo menos um caso de ameaça nessa base, no entanto, uma dessas vítimas registraram 2 ameaças e outras duas registraram 3 ameaças.

Foi observado que os três disparos de arma de fogo ocorreram com a mesma vítima. Quatro boletins foram registrados para essa, além dos três disparos de arma ainda havia 2 lesões corporais, 1 Lesão corporal – violência doméstica e familiar, 1 ameaça, e o feminicídio. Portanto, os disparos de arma estavam registrados nos boletins juntamente com alguma das outras naturezas.

O grupo Contra a honra teve frequência de aproximadamente 4% e Vias de fato 2%. As naturezas relacionadas a esses tiveram baixa frequência e pode se considerar como fator não preponderante para a ocorrência de morte. A injúria foi registrada pelo menos uma vez com 10 vítimas mortas dessa base, o que representa

aproximadamente 7% das linhas. Uma das vítimas teve 2 registros de injúria anterior a sua morte. Dessas 10 vítimas, 8 registraram ameaça, 5, uma vez, uma, 2 vezes e duas 3 vezes. Além das ameaças, 3 das vítimas de injúria sofreram Lesão corporal – violência doméstica e familiar, duas 1 vez e um seis vezes. Também, uma dessas vítimas da injúria sofreu 2 Lesões corporais, outra sofreu 1 vez. A Lesão corporal ocorreu a uma vítima da injúria.

Casos do grupo Violência sexual relacionados a VD nessa contagem representaram aproximadamente 2%. Frequência bem baixa, como nos grupos já citados. Das duas vítimas de Estupro de vulnerável uma também sofreu lesão corporal. Uma vítima com registro da natureza Estupro ou atentado violento ao pudor, essa também registrou 1 Ameaça e 1 Lesão corporal.

As naturezas mais relacionadas ao patrimônio como dano, dano qualificado e furto simples ocorrem em menos de 3% das naturezas. Representando menos de 1% da mesma frequência, apoio a outros órgãos também constam nos boletins dessas vítimas. Essa natureza foi cadastrada juntamente com uma outra natureza criminal, do contrário essa não constaria nessa tabela, o apoio é sem ilicitude e as bases extraídas têm boletins com naturezas criminais. Para que um boletim de ocorrência tenha suas naturezas computadas nessa extração é necessário que pelo menos uma natureza jurídica seja criminal.

Cullen e Fritzson (2019) estudando vítimas fatais de “familiaricídio” (morte entre pessoas da mesma família como cônjuges, ex-cônjuges, pais e filhos) constataram que 34,6% dessas vítimas sofreram ameaça antes da morte. O valor verificado pelos autores na Austrália é 20,6% maior do que o encontrado no Paraná visto no QUADRO 17. Outro fato observado por esses pesquisadores na Austrália é que 36,88% das vítimas tinham histórico de VD. Somando os dados das naturezas jurídicas desde lesão corporal – violência doméstica até Estupro ou atentado violento ao pudor a porcentagem é aproximadamente 57%. Mais da metade das vítimas de feminicídio do Paraná tem registro de naturezas relacionadas a VD, ou seja, tem histórico de VD anterior a sua ocorrência de morte. Então, as vítimas do Paraná têm histórico de VD 20% superior as vítimas estudadas na Austrália.

4.2 RESULTADOS DA MINERAÇÃO DE DADOS COM ALGORITMOS PREDITIVOS DE CLASSIFICAÇÃO

A análise estatística apresentada anteriormente ofereceu fundamentos para determinar os atributos com mais potencial descritivo e que pudessem ser aplicados na etapa de mineração de dados com algoritmos de indução de regras de classificação. Porém, em experimentos de KDD usando técnicas de seleção de atributos com o caso de Souza (2017) concluiu-se que, a amostragem e a seleção de atributos com algoritmos especializados em seleção trazem resultados melhores que a classificação de uma base sem redução de dimensionalidade.

4.2.1 Seleção de atributos das bases

Os atributos das vítimas de feminicídio da base SCOL quando foram submetidos às técnicas de seleção de atributos *CFS*, *Gain Ratio*, *Info Gain*, *ReliefF* ofereceram o município como o atributo mais relevante da base em relação aos atributos meio empregado e motivação. Para o CFS somente o município foi relevante, fato que levou a descontinuar a mineração de dados nessa base.

No caso da base gerada da intersecção entre a contagem das naturezas jurídicas do BOU com as vítimas de feminicídio, as técnicas de seleção de atributos puderam estabelecer as naturezas jurídicas mais relevantes.

Essa base continha 20 atributos: LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR, INJURIA, DESCUMPRIMENTO DE MEDIDA PROTETIVA, AMEACA, DESCUMPRIR DECISÃO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS DE URGÊNCIA PREVISTAS NESTA LEI, boletins, VIAS DE FATO, LESAO CORPORAL DE NATUREZA GRAVISSIMA, CALUNIA, LESAO CORPORAL DE NATUREZA GRAVE, LESAO CORPORAL, PERTURBACAO DA TRANQUILIDADE, DISPARO DE ARMA DE FOGO, FURTO SIMPLES, DANO, DANO QUALIFICADO, CONSTRANGIMENTO ILEGAL, ESTUPRO OU ATENTADO VIOLENTO AO PUDOR, ESTUPRO DE VULNERAVEL e morte.

O atributo morte foi usado como atributo meta para os atributos de seleção, restando então 19 atributos para a redução de dimensionalidade. Iniciando o processo pelo CFS que estabeleceu oito atributos, então foram escolhidos os oito mais bem

ranqueados das outras técnicas. Os atributos selecionados podem ser vistos no QUADRO 18.

QUADRO 18 - ATRIBUTOS SELECIONADOS NA BASE DE NATUREZAS JURÍDICAS DAS VÍTIMAS DE FEMINICÍDIO E VIOLÊNCIA DOMÉSTICA

Técnica	Atributos selecionados
CFS	Lesão corporal Violência doméstica - Ameaça - Lesão corporal de natureza gravíssima - Injúria - Vias de fato - Constrangimento ilegal - Dano Qualificado - Lesão corporal de natureza grave
Info Gain	Injúria - Ameaça - Lesão corporal Violência doméstica - Lesão corporal de natureza gravíssima - Vias de fato - Lesão corporal de natureza grave - boletins - Descumprir Decisão Judicial que defere medidas protetivas de urgência
Gain Ratio	Lesão corporal de natureza gravíssima - Lesão corporal de natureza grave - Dano qualificado - Constrangimento ilegal - Estupro ou atentado violento ao pudor - Disparo de arma de fogo - Injúria - Vias de fato
ReliefF	Injúria - Lesão corporal de natureza gravíssima - Vias de fato - Lesão corporal de natureza grave - Ameaça - Calúnia - Disparo de arma de fogo - Furto simples

FONTE: O Autor (2022).

Foi notado que para as quatro técnicas utilizadas na seleção de atributos, 9 desses foram recorrentes, foram escolhidos pelo menos em duas dessas técnicas. Os atributos: Lesão corporal de natureza grave, Lesão corporal de natureza gravíssima, Vias de fato, Injúria, Ameaça, Constrangimento ilegal, Dano Qualificado, Lesão corporal Violência doméstica, disparo de arma de fogo. Com exceção de disparo de arma de fogo, esses atributos também foram escolhidos pelo algoritmo CFS.

Cada uma das quatro seleções de atributos foi submetida aos dois algoritmos de indução de regras PRISM e CN2. Os resultados da mineração de dados com os algoritmos de indução de regras são apresentados na próxima seção.

4.2.2 Resultados dos algoritmos CN2 e PRISM

Iniciando a etapa de mineração de dados, submeteu-se os quatro conjuntos de atributos selecionados ao algoritmo PRISM. A maior acurácia foi vista na seleção pelas técnicas CFS e Info Gain, 0,548 e melhor área sob a curva pela técnica Info

Gain 0,554. Verificou-se nas matrizes de confusão uma maior capacidade do algoritmo em identificar as vítimas não fatais, acertando em média 98% da classificação para o valor “N”. A técnica Info Gain foi a melhor em classificar os valores da classe morte como “S”, tendo classificado corretamente 16%.

Foi notado que as regras geradas pelo algoritmo eram complexas e com profundidade maior do que 3 em sua maioria. A avaliação objetiva dos resultados gerados pelo algoritmo PRISM inviabilizou o seu envio para avaliação dos especialistas. O fato desse algoritmo não ter gerado um resultado satisfatório nessa etapa, não o inviabiliza para estudos futuros.

Os mesmos quatro conjuntos de atributos gerados pela seleção de atributos foram aplicados ao algoritmo CN2, as medidas de avaliação objetiva podem ser vistas no TABELA 4.

TABELA 4 - RESULTADOS DA EXECUÇÃO DO ALGORITMO CN2 COM AS TÉCNICAS DE SELEÇÃO

Técnicas	Área sob a curva	Acurácia	F1-Score	Precisão	Sensibilidade	Especificidade
CFS	0,701	0,715	0,713	0,724	0,715	0,715
Info Gain	0,689	0,715	0,714	0,720	0,715	0,715
ReliefF	0,663	0,674	0,670	0,680	0,674	0,674
Gain Ratio	0,581	0,615	0,578	0,676	0,615	0,615

FONTE: O Autor (2022).

A técnica de seleção de atributos que gerou os resultados melhores e equilibrados foi a CFS, porém o Info Gain teve valores bem parecidos para acurácia sensibilidade e especificidade. A redução de dimensionalidade realizada pelas técnicas de seleção de atributos gerou um conjunto de regras mais simplificado e significativo.

As matrizes de confusão geradas na execução do algoritmo confirmaram a primeira avaliação feita sob as medidas expostas no QUADRO 19. A escolha feita pelas técnicas CFS e *Info Gain* tiveram os resultados mais equilibrados. Notou-se que esses dois algoritmos erraram menos para as duas classes. Apesar dos algoritmos *ReliefF* e *Gain Ratio* terem apresentados resultados superiores na classificação da classe “S”, esses obtiveram os piores resultados na classe “N”, fato que reforçou a

escolha pelas outras duas técnicas (CFS e *Info Gain*) com resultados mais equilibrados. As matrizes de confusão são demonstradas na FIGURA 3 abaixo.

FIGURA 3 - MATRIZES DE CONFUSÃO GERADAS PELO ALGORITMO CN2

CFS	N	S	Σ	Relief	N	S	Σ
N	117	27	144	N	83	61	144
S	55	89	144	S	33	111	144
Σ	172	116	288	Σ	116	172	288

InfoGain	N	S	Σ	GainRatio	N	S	Σ
N	114	30	144	N	46	98	144
S	52	92	144	S	13	131	144
Σ	166	122	288	Σ	59	229	288

FONTE: O Autor (2022).

O algoritmo CN2 gerou 19 regras com os atributos selecionados pela técnica CFS e 26 para o Info Gain, esses resultados foram submetidos a aprovação dos especialistas do CAPE, um especialista em análise criminal em bases de mortes e outro responsável por análises de VD.

Na análise subjetiva os analistas classificaram cada regra gerada com a letra “A” quando ficou satisfeito com os resultados e surpreso com alguns dos padrões obtidos; letra “B” quando ficou satisfeito com os resultados, mas percebe que já conhecia os padrões obtidos; e letra “C” quando ficou insatisfeito com os resultados. Essa avaliação segue o modelo de avaliação de John (1997) já citado no item 3.4 Avaliação do modelo preditivo.

O algoritmo CN2 com os atributos escolhidos pelas duas técnicas já citadas gerou 10 regras idênticas. As regras com os atributos selecionados pelo CFS foram verificadas primeiro pelos analistas. Das regras criadas, uma recebeu uma letra C de um analista especialista em mortes, regra “DANO QUALIFICADO = 1 - 3” e outra letra C na regra “VIAS DE FATO = 1 - 3” do analista responsável pela VD. A maioria dessas 19 regras receberam A ou B dos dois analistas, 95% delas, sendo A em 74%.

As regras geradas a partir do Info Gain não receberam nenhuma avaliação com letra C, 50% das regras receberam A. Segundo os analistas, foi realizado um estudo sobre a base BOU com vítimas de violência doméstica recentemente no CAPE, alguns conhecimentos gerados naquela análise foram vistos em 13 das regras induzidas com essa técnica de seleção de atributos.

Na análise subjetiva, as regras geradas a partir dos atributos selecionados pelo CFS surpreenderam mais, porém, 13 regras geradas a partir do Info Gain tiveram avaliação com a letra A. Segundo os analistas, as regras induzidas através dos atributos do CFS e do Info Gain acrescentaram conhecimento ao trabalho de análise.

Os dois conjuntos de regras foram aprovados pela avaliação subjetiva, com exceção de duas regras. Excluindo as regras reprovadas, foi possível analisar todas as regras que o algoritmo previu como morte, “S”. Essas regras podem ser vistas agrupadas por natureza jurídica no QUADRO 19.

QUADRO 19 - REGRAS INDUZIDAS COM AS TÉCNICAS DE SELEÇÃO DE ATRIBUTOS CFS E INFO GAIN

REGRAS RESULTANTES DA PREDIÇÃO DO ALGORITMO CN2 COM AS TÉCNICAS CFS E INFO GAIN COM PROBABILIDADE DE MORTE POR REGRA
LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR = 1 - 3 (50%), = 3 - 5 (75%) ou = 7 - 9 (67%)
LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR diferente de 1 - 3 e AMEACA= 0 (74%)
LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR diferente de 1 - 3 e AMEACA diferente de 1 - 3 (74%)
LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR = 0 (50%)
LESAO CORPORAL DE NATUREZA GRAVE = 1 - 3 (83%)
LESAO CORPORAL DE NATUREZA GRAVISSIMA = 1 - 3 (90%)
DESCUMPRIR DECISÃO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS DE URGÊNCIA PREVISTAS NESTA LEI=1 - 3 e LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR diferente de 0 (67%)
DESCUMPRIR DECISÃO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS DE URGÊNCIA PREVISTAS NESTA LEI=1 - 3 e AMEACA diferente de 1 - 3 (75%)
AMEACA = 0 e boletins diferente de 1 - 3 (83%)
CONSTRANGIMENTO ILEGAL = 1 - 3 (67%)
VIAS DE FATO = 1 - 3 e boletins = 1 - 3 (71%)

FONTE: O Autor (2022).

No conjunto de regras que previu a morte, foi adicionada a probabilidade que o algoritmo calculou por regra. As regras com probabilidades acima de 60% têm uma frequência de maior que as abaixo dessa. Portanto, com exceção de uma das regras,

as probabilidades sugerem uma descoberta consistente sobre a previsão para as mortes. Assim, este quadro representa o perfil de ocorrências previsto para uma vítima de VD com propensão a morte.

Nota-se que as Lesões corporais e ameaça estiveram em 7 dessas 11 regras, sugerindo que o acompanhamento dessas naturezas jurídicas nos casos de mulheres vítimas de VD é muito importante para prevenir a morte.

4.3 FLUXO DE ANÁLISE PREDITIVA DE DADOS USANDO KDD EM DADOS DE VÍTIMAS FATAIS

A criação de um fluxo de análise preditiva de dados proposta neste estudo foi pensado para apoiar pesquisas em dados de violência doméstica e feminicídio usando o método KDD. Porém, pesquisas futuras utilizando bases de dados de VD contra a mulher podem ter objetivos diversos dos desta análise. Além disso, bases de dados fornecidas por órgãos de segurança pública, dependendo do estado, podem ter diferentes apresentações, formatos e atributos.

Entendeu-se que um único fluxo não poderia apoiar tantas possibilidades de pesquisa na área de VD. Então criou-se três fluxos de análise de dados: um com estatística descritiva, um com a utilização do KDD e um com KDD para bases de dados com reincidências.

O primeiro fluxo, denominado “**fluxo de análise de dados preditiva com estatística descritiva**”, é indicado para pesquisas que têm por objetivo criar um perfil de vítima de VD utilizando a análise de frequências de atributos.

Neste fluxo, é realizado um pré-processamento e um tratamento nos dados. Em muitas bases de dados de VD extraídas de sistemas da segurança pública, pode existir algum problema de completude de atributos não obrigatórios, atributos redundantes e problemas na padronização de valores de atributos abertos. Para tanto, foram previstos procedimentos antes da análise de frequências dos atributos. No pré-processamento deste fluxo, são propostos procedimentos de exclusão de atributos com completude abaixo de 60%, exclusão de atributos redundantes e padronização de valores. O fluxograma desse fluxo pode ser observado na FIGURA 4 abaixo.

O segundo fluxo, denominado “**fluxo de análise de dados preditiva com KDD**”, é indicado para pesquisas que tenham objetivos de criar perfil de vítima de VD contra a mulher utilizando o KDD. A base SCOL é um exemplo de base de dados que pode ser aplicada neste fluxo. Na mesma linha estão atributos que caracterizam uma vítima de feminicídio, o local, a data, a hora e as características do crime (meio empregado, motivação etc.).

O processo inicial deste fluxo é o carregamento da base pré-processada e tratada no primeiro fluxo. Portanto, os atributos redundantes ou com completude abaixo de 60% não devem fazer parte da base de dados aplicada nele. Ainda, os valores de atributos abertos devem estar padronizados, procedimentos propostos no primeiro fluxo de análise.

Na etapa de pré-processamento deste fluxo, é proposto a escolha de um atributo “meta”. Algoritmos de classificação como os de indução de regras necessitam deste atributo definido para sua execução. Sugere-se uma revisão nos objetivos da pesquisa antes da escolha do atributo ou atributos meta utilizando este fluxo.

Tendo selecionado o atributo meta, este fluxo propõe a verificação do balanceamento desse atributo meta. Em um exemplo, na base SCOL, a estatística descritiva explicando o atributo “autor conhecido” evidenciou um grande desbalanceamento, em torno de 98% dos crimes de feminicídio no Paraná tem o seu autor conhecido. Se hipoteticamente o objetivo da pesquisa fosse explicar a autoria em crimes de feminicídio esse atributo poderia ser o “meta” em alguma análise.

O atributo de autoria já citado tem “S” ou “N” como valores possíveis, se 98% das linhas contêm o valor “S” e 2% contêm o valor “N”, a distribuição é muito desbalanceada, muitos dos algoritmos de classificação não são sensíveis ao desbalanceamento dificultando o processo de predição. Portanto, sugere-se o balanceamento de classes com a função “*uppersample*” para atributos com classes desbalanceadas. Esta função está implementada no pacote “*Themis*” do RStudio.

Ainda no pré-processamento, um procedimento importante trata da escolha de atributos mais relevantes da base. Quatro técnicas com a abordagem filtro são propostas nessa etapa: CFS, Gain Ratio, Info Gain e ReliefF. Todos foram explicados na revisão de literatura. Sugere-se que o processo seja iniciado no software WEKA com a técnica CFS. A partir do resultado do CFS, utilize o software Orange carregando a mesma base para as outras três técnicas.

O processo de seleção de atributos é importante neste fluxo de KDD, é muito difícil prever a quantidade de atributos em uma base com dados de VD. Uma base pode conter muitos atributos que não são relevantes no processo de predição, as técnicas diminuem a dimensionalidade da base e tendem a melhorar os resultados.

O processo de *data mining* proposto oferece algumas sugestões de algoritmos que podem ser utilizados nessa etapa do KDD. Algoritmos como o PRISM, CN2 e PART estão implementados no software WEKA e Orange, dependendo da base de dados e objetivo da análise, um algoritmo pode ser mais preditivo que outro. É importante utilizar mais de um algoritmo nessa etapa, pois, por mais que se entenda o potencial preditivo de um algoritmo, os resultados ainda dependem do tipo de base de dados utilizada na análise.

Os resultados das execuções dos algoritmos com os atributos selecionados pelas técnicas devem ser tabulados. Cada linha da tabela deve conter o nome do algoritmo e da técnica de seleção separados por uma barra, identificando a que se referem as medidas de avaliação anotadas.

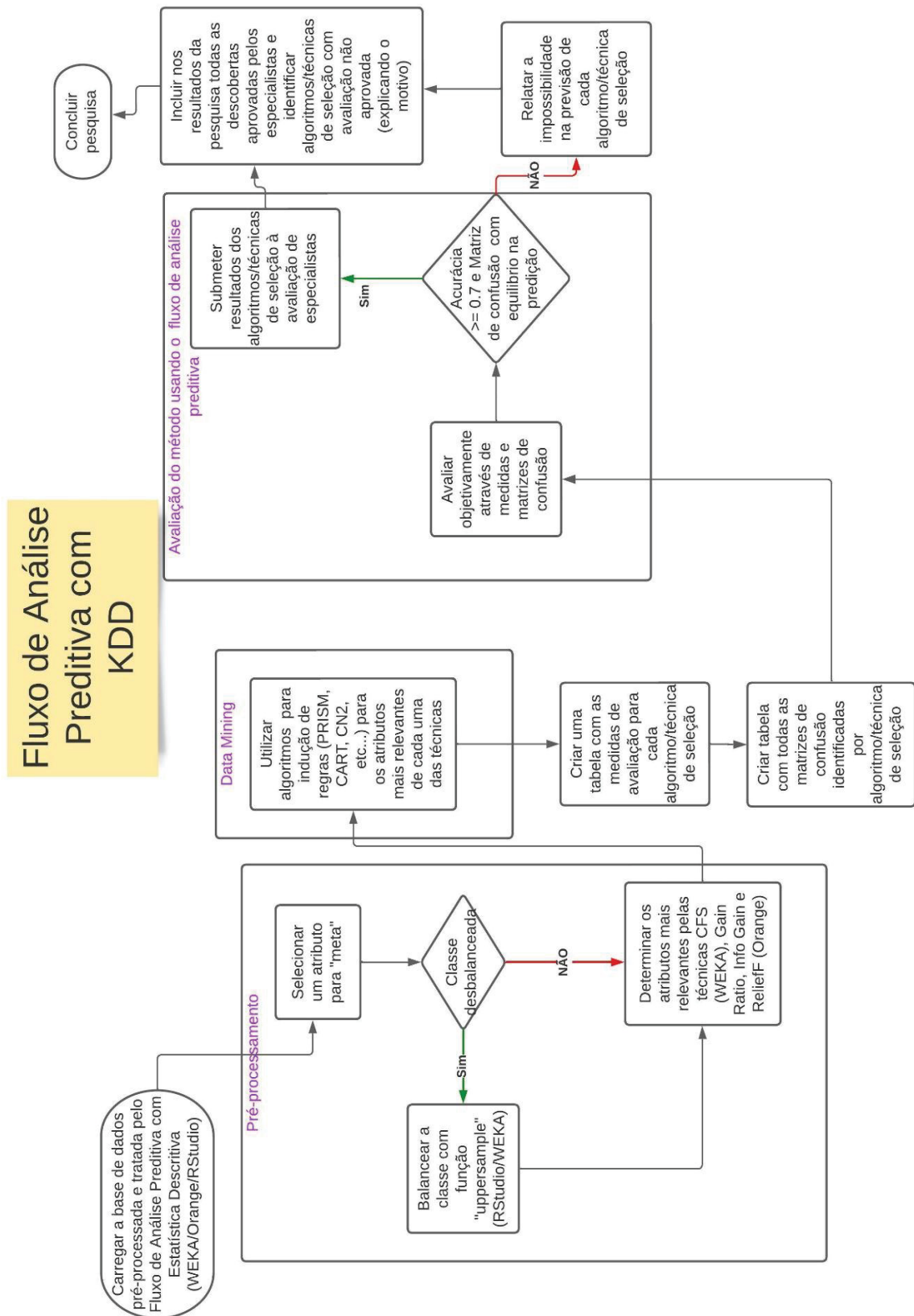
Com as medidas tabuladas por algoritmo/técnica de seleção, é necessário a tabulação das matrizes de confusão com a mesma indicação de algoritmo e técnica de seleção.

Tendo tabuladas as medidas e matrizes, o próximo processo trata da avaliação das medidas e matrizes de confusão objetivamente. Sugere-se que se aprove para o próximo processo a execução que tenha a acurácia de ao menos 0,7 e as predições para as classes tenham equilíbrio na matriz de confusão. Portanto, mesmo que a acurácia seja maior que 0,7, se as predições errarem muito mais em uma classe, os resultados desse algoritmo/técnica de seleção não devem ser enviados para aprovação de especialistas.

Todos os resultados aprovados pela avaliação objetiva (medidas e matrizes de confusão), segundo o fluxo proposto, devem ser avaliados por especialistas. Isso, para garantir que o resultado aprovado tem utilidade. O modelo de avaliação sugerido é o de John (1997), já explicado no item 3.4.

Por fim, todos os resultados aprovados são adicionados aos resultados de forma sumarizada. Se a pesquisa utilizar algoritmos de indução de regras, é importante descrever como se deu a aprovação das regras, quais foram melhor/pior avaliadas. Também é importante relatar quais resultados não foram aprovados pela avaliação objetiva. A FIGURA 5 ilustra o segundo fluxo abaixo.

FIGURA 5 - FLUXOGRAMA DE ANÁLISE DE DADOS PREDITIVA COM KDD



FONTE: O Autor (2022).

Idealizou-se o terceiro fluxo para tornar possível a análise de reincidências de crimes com uma mesma vítima de VD. Este foi denominado **“fluxo de análise de dados preditiva com KDD para base com reincidências”**.

Em uma base de dados de VD, a exemplo do BOU, cada linha tem dados da vítima, características dessa, local, data, hora, natureza jurídica do crime, meio empregado, entre outros atributos. Se para uma mesma ocorrência forem cadastradas duas naturezas jurídicas diferentes, ocorre a duplicação de linhas na extração.

A natureza jurídica é utilizada para identificar a ação criminosa da ocorrência, ou seja, esta explica muito do ocorrido em poucas palavras. Então, propõe-se no fluxo o desmembramento desse atributo, agrupando sua contagem por vítima.

O agrupamento de naturezas por vítima pode ser feito em partes, extraindo do banco de dados a contagem com o comando “group by” (a exemplo da base Naturezas) ou por softwares como o RStudio, realizando a mesma contagem em bases já extraídas.

Após a primeira etapa de agrupamento, é preciso alinhar a vítima com as naturezas e contagens na mesma linha. Recomenda-se o uso da função *pivot_wider* do pacote *dplyr* do RStudio. Ao fim dessa etapa de pré-processamento, tem-se uma base de dados com uma vítima, suas naturezas e as contagens destas.

No tratamento, o próximo processo propõe a criação de um atributo denominado “morte”. Depois, atribuir o valor “S” para vítimas mortas e “N” para não mortas. Para realizar essa atribuição, a sugestão é usar um teste lógico com as naturezas relacionadas à morte, como feminicídio, homicídio, atendimento do local de morte, etc. Como o nome das naturezas jurídicas pode apresentar pequenas diferenças de estado para estado, é importante consultar especialistas para determinar quais são as naturezas que representam a morte da vítima. Esse tratamento, criar um atributo e atribuir um valor baseado na natureza, pode ser realizado no RStudio com o pacote *dplyr*.

O próximo processo propõe selecionar somente naturezas jurídicas relacionadas às vítimas mortas para os próximos procedimentos. Isso pois, pretende-se comparar as vítimas mortas com não mortas, então, mantém-se na base o que é igual para vítimas mortas e não mortas.

Os próximos processos deste fluxo são idênticos ao do anterior a partir do processo “selecionar o atributo morte para meta”. A diferença deles está no pré-

5 CONSIDERAÇÕES FINAIS

A análise de dados de VD no Brasil depende da autorização de órgãos governamentais, portanto, no planejamento de uma pesquisa acadêmica com a pretensão de utilizar dados como esses, a autorização deve ser colocada como o mais importante. A autorização do uso dos dados em pesquisa concedida pela SESP/PR a este pesquisador e orientadora não tem precedentes. Sendo assim, a oportunidade de realizar uma pesquisa dessa profundidade em dados protegidos por leis e decretos é um fato raro.

Na descrição do problema de pesquisa da primeira seção, (Como determinar quais atributos devem ser considerados na proposta de um fluxo de análise preditiva para identificação vítimas fatais de violência doméstica?), na execução das quatro técnicas de seleção de atributos já citadas nos resultados, na base SCOL, foi verificado que somente o atributo município foi escolhido como atributo importante em relação ao meio empregado e motivação do crime.

No entanto, quando foram aplicadas as técnicas de seleção de atributos (*CFS*, *Gain Ratio*, *Info Gain*, *ReliefF*) na base de ocorrências com vítimas fatais e não fatais verificou-se uma consistência na escolha das naturezas jurídicas: Ameaça, Constrangimento ilegal, Dano qualificado, Disparo de arma de fogo, Injúria, Lesão corporal de natureza grave, Lesão corporal de natureza gravíssima, Lesão corporal Violência doméstica, Vias de fato. As técnicas de seleção *CFS* e *Info Gain* produziram os melhores resultados para essa base. As regras geradas pelo algoritmo CN2 com a seleção de atributos foi mais enxuta, assertiva e com maior qualidade se comparada as regras geradas com todos os 19 atributos, a base sem redução de dimensionalidade.

O objetivo específico a, (Analisar e descrever o processo atual de recuperação de informação e produção de relatórios sobre o tema Violência Doméstica no Centro de Análise, Planejamento e Estatística) foi realizado na seção 3 com destaque para a novidade importante para pesquisadores, o acesso público concedido pelo estado a alguns dados de VD.

O objetivo específico b, (Aplicar a Descoberta de Conhecimento em Bases de Dados para encontrar padrões nos dados de violência doméstica quando a vítima foi morta), foi realizado na seção 4 e teve êxito em extrair padrões desconhecidos de atributos de naturezas jurídicas das ocorrências. A técnica de amostragem

undersampler e as técnicas de seleção de atributos *CFS*, *Gain Ratio*, *Info Gain*, *ReliefF*, tornaram a descoberta possível, simplificando o universo das naturezas jurídicas de ocorrências de violência doméstica no Paraná nos anos estudados.

O objetivo específico c, (Analisar e comparar os resultados dos métodos de Indução de regras), foi observado que o algoritmo CN2 pode gerar mais regras com padrões desconhecidos, a forma de apresentação do relatório é mais prática na apresentação dos resultados, a indexação do método é ordenada pela qualidade da regra e apresenta as probabilidades por regra criada. Então, dos algoritmos utilizados na análise, o CN2 teve melhor desempenho. Ainda, os resultados gerados pelo algoritmo PRISM foram reprovados na avaliação objetiva pelo desequilíbrio nas medidas e ser ineficiente ao prever a classe “S”, que representa a morte.

Quanto ao objetivo geral (Propor um fluxo de análise de dados preditiva que possibilite determinar um perfil para as mulheres vítimas fatais de violência doméstica), três fluxos foram propostos nos resultados. É difícil prever quais serão objetivos de novas pesquisas nessa área ou ter certeza de como serão as bases de dados de dados utilizadas nas análises. Então, esses fluxos foram criados para oferecer um caminho para novas pesquisas de análise de dados em VD, descrevendo pelo menos três formas de se analisar. Portanto, esses fluxos contribuem tanto para pesquisa acadêmica quanto para pesquisas realizadas por especialistas em análise de dados da segurança pública.

Na revisão integrativa não foi encontrada nenhuma ferramenta de apoio à pesquisa em dados de VD como esses fluxos. Então, entende-se que os três fluxogramas expostos nos resultados podem ser utilizados por pesquisadores como referência para novas análises de dados de VD com estatística descritiva e KDD.

Segundo os analistas do CAPE, a ligação dos dados das bases Naturezas e SCOL quando se faz necessária para análise por demanda leva bastante tempo e é realizada de forma não automatizada. Algumas regras geradas pelo algoritmo CN2 foram surpreendentes para os especialistas pois trouxeram tanto conhecimentos sobre a VD que esses assimilaram durante sua carreira quanto evidenciaram características dessa base de dados ainda desconhecidas por eles.

As bases de dados extraídas do sistema BOU (base BOU e Naturezas) necessitaram de um grande esforço no pré-processamento, limpeza e tratamento para possibilitar seu uso na estatística descritiva e KDD. O pré-processamento e o tratamento exigem um tempo maior de realização comparado às outras etapas de

KDD, fato constatado nessa análise, entretanto, a criação de scripts no RStudio dessa análise permite sua reutilização tanto para uma atualização da pesquisa quanto a automatização de alguns procedimentos realizados nas bases.

Portanto, sugere-se rever a infraestrutura da base de dados do sistema BOU com foco na recuperação de dados. O modelo atual de recuperação de dados de sistemas como o BOU não foi idealizado para extrair bases de dados próprias para uso em análises com KDD, o que representa um impedimento na automatização de análises de dados como as de VD que tendem a ocupar muito tempo dos analistas.

A mineração de dados realizada sobre uma amostra com 20 atributos de 288 vítimas da base de dados unificada (SCOL e BOU) teve êxito em descobrir conhecimento novo e útil. As técnicas de seleção de atributos CFS e Info Gain tiveram os resultados mais equilibrados e de fato reduziram a dimensionalidade da base gerando conhecimento novo e útil no algoritmo CN2.

O resultado das análises pode determinar que as vítimas de feminicídio no Paraná, dos anos 2018, 2019 e 2020, sofreram mais de violência física do que de outros tipos de violência. A violência psicológica (ameaça) é o segundo tipo de violência mais frequente nos dados. O meio empregado na morte mais comum é arma branca ou imprópria, seguido por disparo de arma de fogo e agressão física. A autoria dos crimes é conhecida na grande maioria das vezes. As mortes são mais frequentes nos finais de semana e segunda-feira. Vítimas com 29 anos foram observadas 14 vezes, maior frequência para uma idade na base SCOL. Para pesquisas futuras, propõe-se: avaliar a capacidade preditiva das redes neurais artificiais em uma base de dados de naturezas jurídicas e atributos demográficos agrupados por vítima; submeter os dados com e sem tratamento à mineração de processos avaliar as descobertas e propor melhorias.

REFERÊNCIAS

AGHAKHANI, N. *et al.* Study of the types of domestic violence committed against women referred to the legal medical organization in Urmia - Iran. **Iranian Journal of Psychiatry and Behavioral Sciences**, [s. l.], v. 9, n. 4, p. 1-6, 2015. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4733310/pdf/ijpbs-09-2446.pdf>. Acesso em: 12 mar. 2021.

AZEVEDO, A.; SANTOS, M. F. KDD, SEMMA AND CRISP-DM: a parallel overview. In: IADIS European Conference on Data Mining, [s. l.], 2008, Amsterdam. **Proceedings...** Amsterdam: IADIS, 2008. p. 182–185. Disponível em: <http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf>. Acesso em: 10 out. 2021.

BEAL, A. **Gestão Estratégica da Informação**: como transformar a informação e a tecnologia da informação em fatores de crescimento e de alto desempenho nas organizações. 3. ed. São Paulo: Atlas, 2008.

BORGES, H. B. **Redução de dimensionalidade em bases de dados de expressão gênica**. 2006. 123 f. Dissertação (Mestrado em Informática) - Programa de Pós-graduação em Informática, Pontifícia Universidade Católica do Paraná, Curitiba, 2006. Disponível em: https://www.ppgia.pucpr.br/pt/arquivos/mestrado/dissertacoes/2006/2006_helyane.pdf. Acesso em: 02 fev. 2022.

BRASIL. Decreto-Lei 2.848, de 07 de dezembro de 1940. Código Penal. **Diário Oficial da União**, Brasília, DF, 31 dez. 1940. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848.htm. Acesso em: 16 maio 2022.

BRASIL. Lei 13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). **Diário Oficial da União**, Brasília, DF, n. 158, 14 ago. 2018. Seção 1, p. 59. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/36849373/do1-2018-08-15-lei-no-13-709-de-14-de-agosto-de-2018-36849337. Acesso em: 16 maio 2022.

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. **Diário Oficial da União**, Brasília, DF, n. 151, 8 ago. 2006. Seção 1, p. 1. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm. Acesso em: 16 maio 2022.

BRASIL. Lei nº 13.104, de 9 de março de 2015. Altera o art. 121 do Decreto-Lei nº 2.848, de 7 de dezembro de 1940 - Código Penal, para prever o feminicídio como circunstância qualificadora do crime de homicídio, e o art. 1º da Lei nº 8.072, de 25 de julho de 1990, para incluir o feminicídio no rol dos crimes hediondos. **Diário Oficial da União**, Brasília, DF, n. 46, 10 mar. 2015. Seção 1, p. 1. Disponível em: <https://www2.camara.leg.br/legin/fed/lei/2015/lei-13104-9-marco-2015-780225-publicacaooriginal-146279-pl.html>. Acesso em: 16 maio 2022.

BRASIL. Lei nº 14.188, de 28 de julho de 2021. Define o programa de cooperação Sinal Vermelho contra a Violência Doméstica como uma das medidas de enfrentamento da violência doméstica e familiar contra a mulher previstas na Lei nº 11.340, de 7 de agosto de 2006 (Lei Maria da Penha), e no Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), em todo o território nacional; e altera o Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), para modificar a modalidade da pena da lesão corporal simples cometida contra a mulher por razões da condição do sexo feminino e para criar o tipo penal de violência psicológica contra a mulher. **Diário Oficial da União**, Brasília, DF, n.142, 29 jul. 2021. Seção 1, p. 1. Disponível em: <https://www.in.gov.br/en/web/dou/-/lei-n-14.188-de-28-de-julho-de-2021-334902612>. Acesso em: 16 maio 2022.

CAPRARA, G. L. *et al.* Does domestic violence during pregnancy influence the beginning of complementary feeding? **BMC Pregnancy and Childbirth**, Campus, 4 Crinan St, London, England, v. 20, n. 1, 2020. Disponível em: <https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-020-03144-y>. Acesso em: 20 abr. 2021.

CENDROWSKA, J. **PRISM**: An algorithm for inducing modular rules. [S. l.: s. n.], 1987. Disponível em: <https://sci2s.ugr.es/keel/pdf/algorithm/articulo/1987-Cendrowska-IJMMS.pdf>. Acesso em: 05 mar. 2022.

CHAWLA, N. V. **Data mining for imbalanced datasets**: an overview. In: Data Mining and Knowledge Discovery Handbook. MAIMON, In: Oded; ROKACH, Lior (org.). [S. l.]: Springer, 2005. p. 853–867. Disponível em: https://link.springer.com/chapter/10.1007/0-387-25465-X_40. Acesso em: 15 maio 2022.

CLARK, P.; NIBLETT, T. **The CN2 induction algorithm machine learning**. [S. l.: s. n.], 1989. Disponível em: <https://link.springer.com/content/pdf/10.1007/BF00116835.pdf>. Acesso em: 18 abr. 2022.

CLARKE, A. *et al.* Violence exposure and young people's vulnerability, mental and physical health. **International Journal of Public Health**, [s. l.], v. 65, n. 3, p. 357–366, 2020. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7183502/>. Acesso em: 20 abr. 2021.

CÔRTEZ, S. Da C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados – funcionalidades, técnicas e abordagens. **PUC-Rio Informática**, [s. l.], p. 35, 2002. Disponível em: <https://docplayer.com.br/694226-Mineracao-de-dados-funcionalidades-tecnicas-e-abordagens.html>. Acesso em: 15 maio 2022.

CRESPO, A. A. **Estatística Fácil**. 17. ed. São Paulo: Saraiva, 2002.

CULLEN, D.; FRITZON, K. A typology of familicide perpetrators in Australia. **Psychiatry, Psychology and Law**, [s. l.], v. 26, n. 6, p. 970–988, 2019. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7033703/https://www.tandfonline.com/doi/full/10.1080/13218719.2019.1664276>. Acesso em: 12 abr. 2021.

DA SILVA COSTA, C. A.; TSUNODA, D. F.; PECINI, A. C. Análise de dados de violência doméstica: revisão integrativa. In: CONGRESSO NACIONAL DE ADMINISTRAÇÃO, 26., 2021, Goiânia. **Anais...** Goiânia: SINAGO, 2021. p. 65-76. Disponível em: https://drive.google.com/file/d/1R-y8s_mEP6LrelU9NW2OQq2jnoSWTeL/view. Acesso em: 20 jul. 2022.

DA SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicação em R. 1. ed. Rio de Janeiro: Elsevier, 2016.

DASGUPTA, A. *et al.* Local social support mitigates depression among women contending with spousal violence and husband's risky drinking in Mumbai slum communities. **Journal of Affective Disorders**, [s. l.], v. 145, n. 1, p. 126–129, 2013. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S016503271200403X?via%3Di> hub. Acesso em: 12 abr. 2021.

DAVENPORT, T. H. **Ecologia da informação**: por que só a tecnologia não basta para o sucesso na era da Informação. São Paulo: Futura, 1998.

DE ANDRADE, M. D. A utilização do sistema RStudio e da jurimetria como ferramentas complementares à pesquisa jurídica. **REVISTA QUAESTIO IURIS**, [s. l.], v. 11, n. 2, p. 680-692, 2018. Disponível em: <https://www.e-publicacoes.uerj.br/index.php/quaestioiuris/article/view/29221>. Acesso em: 17 jun. 2022.

DERRICK, J. L.; TESTA, M.; LEONARD, K. E. Daily reports of intimate partner verbal aggression by self and partner: Short-term consequences and implications for measurement. **Psychology of Violence**, [s. l.], v. 4, n. 4, p. 416–431, 2014. Disponível em: <https://psycnet.apa.org/record/2014-33180-001>. Acesso em 12 abr. 2021.

DIEHL, C. *et al.* O uso da estatística descritiva na pesquisa em custos: Análise do XIV Congresso Brasileiro De Custos. **ConTexto**, [s. l.], v. 7, n. 12, p. 1-24, 2007. Disponível em: <https://seer.ufrgs.br/index.php/ConTexto/article/view/11157/6605>. Acesso em 18 jun. 2022.

DUTTA, T. *et al.* Association between individual and intimate partner factors and cervical cancer screening in Kenya. **Preventing Chronic Disease**, 1600 Clifton RD, Atlanta, GA, USA, v. 15, n. 12, p. 1-9, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0->

85058919746&doi=10.5888%2Fpcd15.180182&partnerID=40&md5=99f65d115bd6255aef470d73c1150c55. Acesso em: 12 abr. 2021.

ENRIQUEZ-CANTO, Y. *et al.* Ecological analysis of intimate partner sexual violence in peruvian women. **ACTA COLOMBIANA DE PSICOLOGIA**, Carrera 13 N° 47-49, Bogota D.C., Colombia, v. 23, n. 1, p. 287–300, 2020. Disponível em: http://www.scielo.org.co/pdf/acp/v23n1/es_0123-9155-acp-23-01-272.pdf. Acesso em: 12 abr. 2021.

ERCOLE, F. F.; MELO, L. S. de; ALCOFORADO, C. L. G. C. Integrative review versus systematic review. **Reme: Revista Mineira de Enfermagem**, [s. l.], v. 18, n. 1, p. 9-11, 2014. Disponível em: <http://www.revenf.bvs.br/pdf/remex/v18n1/v18n1a01.pdf>. Acesso em: 21 maio 2021.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. Knowledge discovery and data mining. **Association for the Advancement of Artificial Intelligence (AAAI)**, [s. l.], v. 17, n. 3, p. 37–54, 1996. Disponível em: <https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf>. Acesso em: 05 fev. 2022.

FREIXO, M. J. V. **Metodologia científica: fundamentos métodos e técnicas**. 2. ed. São Paulo: Instituto Piaget, 2010.

FULLER-THOMSON, E.; RAMZAN, N.; BAIRD, S. L. Arthritis and suicide attempts: findings from a large nationally representative Canadian survey. **Rheumatology International**, [s. l.], v. 36, n. 9, p. 1237–1248, 2016.

GALVÃO, M. C. B.; RICARTE, I. L. M. Revisão sistemática da literatura: conceituação, produção e publicação. **Logeion: Filosofia da Informação**, [s. l.], v. 6, n. 1, p. 57–73, 2019. Disponível em: <https://revista.ibict.br/fiinf/article/view/4835/4187>. Acesso em: 12 abr. 2022.

GEBREZGI, B.H. *et al.* Factors associated with intimate partner physical violence among women attending antenatal care in Shire Endasselassie town, Tigray, northern Ethiopia: A cross-sectional study, July 2015. **Reproductive Health**, [s. l.], v. 14, n. 1, 2017. Disponível em: <https://reproductive-health-journal.biomedcentral.com/articles/10.1186/s12978-017-0337-y>. Acesso em: 12 abr. 2022.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008. ISSN 00034886.v. 10.

GOLDSHMIDT, R. O.; PASSOS, E. **Data mining: conceitos, técnicas, algoritmos, orientações e aplicações**. Rio de Janeiro: Elsevier, 2005. ISSN 07102755.

HALL, M. A. **Correlation-based feature selection for discrete and numeric class machine learning**. Hamilton, New Zealand: [s. n.], 2000. Disponível em: <https://researchcommons.waikato.ac.nz/handle/10289/1024>. Acesso em: 10 ago. 2022.

HALMENSCHLAGER, C. **Um algoritmo para indução de árvores e regras de decisão**. Orientador: Luís Otavio Campos Alvares. 2002. 112 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/2755/000325797.pdf>. Acesso em: 08 ago. 2022.

HAN, Y-R.; CHOI, H.Y. Risk factors affecting intimate partner violence occurrence in South Korea: findings from the 2016 domestic violence survey. **PLOS ONE**, [s. l.], v. 16, p. 1-16, 3 March, 2021. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0247916>. Acesso em: 12 abr. 2021.

JOVIĆ, A; BRKIĆ, K; BOGUNOVIĆ, N. An overview of free software tools for general data mining. In: International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 37., 2014. **Proceedings...** Opatija, Croácia: MIPRO, 2014. p. 1112–1117. Disponível em: https://www.researchgate.net/publication/269291367_An_overview_of_free_software_tools_for_general_data_mining. Acesso em: 22 mar. 2022.

KARYSTIANIS, G. *et al.* Automated analysis of domestic violence police reports to explore abuse types and victim injuries: text mining study. **Journal of Medical Internet Research**, [s. l.], v. 21, n. 3, p. 1-12, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062834749&doi=10.2196%2F13067&partnerID=40&md5=d5064fbad57a16a0acf7f7736faaf01>. Acesso em: 12 abr. 2021.

KHAYAT, S. *et al.* Association between physical and sexual violence and mental health in suburban women of zahedan: a cross-sectional study. **Journal of Clinical and Diagnostic Research**, [s. l.], v. 11, n. 12, p. IC01–IC05, 2017. Disponível em: https://www.researchgate.net/publication/322271055_Association_between_Physical_and_Sexual_Violence_and_Mental_Health_in_Suburban_Women_of_Zahedan_A_Cross-sectional_Study/link/5a52fc9aa6fdcc7690042327/download. Acesso em: 12 abr. 2021.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. [S. l.: s. n.], 2007. Disponível em: https://www.elsevier.com/___data/promis_misc/525444systematicreviewsguide.pdf. Acesso em: 12 maio 2022.

LAEHEEM, K.; BOONPRAKARN, K. Family background risk factors associated with domestic violence among married thai muslims couples in Pattani Province. **Asian Social Science**, [s. l.], v. 11, n. 9, p. 235–244, 2015. Disponível em: <https://ccsenet.org/journal/index.php/ass/article/view/47165>. Acesso em: 12 abr. 2021.

LEE, P. H. Resampling methods improve the predictive power of modeling in class-imbalanced datasets. **International Journal of Environmental Research and Public Health**, [s. l.], v. 11, n. 9, p. 9776–9789, 2014. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4199049/>. Acesso em: 27 jul. 2022.

LEE, H. D. **Seleção de atributos importantes para a extração de conhecimento de bases de dados**. Orientadora: Maria Carolina Monard. 2005. 154 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2005. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/publico/tese_huei.pdf. Acesso em: 10 ago. 2022.

LIMA, R. A. F. de. **Estratégias de seleção de atributos para detecção de anomalias em transações eletrônicas**. Orientador: Adriano César Machado Pereira. 2016. 94 f. Dissertação (Mestrado em Ciência da Computação) – Departamento de Ciência da Computação, Universidade Federal de Minas Gerais - UFMG, Belo Horizonte, 2016. Disponível em: <https://www.dcc.ufmg.br/pos/cursos/defesas/1930M.PDF>. Acesso em: 09 ago. 2022.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, [s. l.], v. 17, n. 4, p. 491–502, 2005. Disponível em: <https://sci2s.ugr.es/keel/pdf/algorithm/articulo/2005-Liu-IEEETKDE.pdf>. Acesso em: 15 ago. 2022.

MANZANARES, R. C.; TORRADO TARRÍO, C.; SALGADO, C. A. Víctima y derecho penal mediación em violência de gênero. **Revista de Mediación**. [S. l.: s. n.], 2011. Disponível em: <https://revistademediacion.com/wp-content/uploads/2013/10/Revista-Mediacion-7-05.pdf>. Acesso em: 7 mar. 2022.

MARCONI, M.; LAKATOS, E. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

MARIANO, D. Métricas de Avaliação em Machine Learning. **BIOINFO**, v. 1, p. 233–241, jul. 2021. Disponível em: <https://bioinfo.com.br/metricas-de-avaliacao-em-machine-learning-acuracia-sensibilidade-precisao-especificidade-e-f-score/>. Acesso em: 12 ago. 2022.

MARTINS, G. de A.; THEÓPHILO, C. R. P. **Metodologia da investigação científica para ciências sociais aplicadas**. São Paulo: Atlas, 2009.

MCGEE, J. V.; PRUSAK, L. **Gerenciamento estratégico da informação: aumente a competitividade e a eficácia de sua empresa utilizando a informação como uma ferramenta estratégica**. Rio de Janeiro: Campus, 1994.

MELO, R. Teoria crítica e os sentidos da emancipação. **Caderno CRH**, [s. l.], v. 24, n. 62, p. 249–262, 2011. Disponível em: <https://www.redalyc.org/articulo.oa?id=347632184002>. Acesso em: 18 set. 2022.

NEVES, R. de C. D. das. **Pré-processamento no processo de descoberta de conhecimento em banco de dados**. Orientador: Luís Otavio Campos Alvares. 2003. 137 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003. Disponível em:

<https://www.lume.ufrgs.br/bitstream/handle/10183/2701/000375412.pdf>. Acesso em: 12 abr. 2022.

NOBRE, M. **A teoria crítica**. 1. ed. Rio de Janeiro: Zahar, 2004.

OMS. **World report on violence** (World Health Organization, Org.). Geneva: [s. n.], 2002. Disponível em: http://apps.who.int/iris/bitstream/handle/10665/42495/9241545615_eng.pdf;jsessionid=137A4A01754A03065E3C68CD7262D858?sequence=1. Acesso em: 16 mar. 2021.

QUINLAN, J. R. Simplifying decision trees. **International Journal of Man-Machine Studies**, [s. l.], v. 27, n. 3, p. 221–234, 1987. Disponível em: Acesso em: 19 jul. 2022.

SILWAL, A; THAPA, B. Prevalence of domestic violence among infertile women attending subfertility clinic of a tertiary hospital. **Journal of the Nepal Medical Association**, [s. l.], v. 58, n. 226, p. 372–376, 2020. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087834552&doi=10.31729%2Fjnma.4886&partnerID=40&md5=9f90708b0ebb0e0914fdcd4b4d0db89e>. Acesso em: 12 abr. 2021.

SOUTO, R.Q. *et al.* Older Brazilian women's experience of psychological domestic violence: a social phenomenological study. **International Journal for Equity in Health**, [s. l.], v. 14, n. 1, p. 1-9, 2015. Disponível em: <https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-015-0173-z>. Acesso em: 12 abr. 2021.

SOUZA, J. T. de. **Métodos de seleção de atributos e análise de componentes principais**: um estudo comparativo. Orientador: Antonio Carlos de Francisco. 2017. 73 f. Dissertação (Mestrado em Engenharia da Produção) - Universidade Tecnológica Federal do Paraná, Programa de Pós-Graduação em Engenharia de Produção, Ponta Grossa, 2017. Disponível em: https://repositorio.utfpr.edu.br/jspui/bitstream/1/2387/1/PG_PPGEP_M_Souza%2C%20Jovani%20Taveira%20de_2017.pdf. Acesso em: 17 ago. 2022.

TANRIVERDI, G; ÇAPIK, C; YALÇIN GÜRSOY, M. Prevalence of domestic violence against married women in Turkey and associated risk factors. **Türkiye Klinikleri Journal of Medical Sciences**, Department of Nursing, Çanakkale Onsekiz Mart University, School of Health, Çanakkale, Turkey, v. 38, n. 3, p. 218–229, 2018. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053329565&doi=10.5336%2Fmedsci.2017-58822&partnerID=40&md5=f1e7b858d7ce0ff5c51a8d2c2a8c1c75>. Acesso em: 12 abr. 2022.

VASCONCELOS, B. de S. **Mineração de regras de classificação com sistemas de banco de dados objeto-relacional**: Estudo de caso: classificação de litofácies de poços de petróleo. Orientador: Marcus Costa Sampaio. 2002. 127 f. Dissertação (Mestrado em Ciência da Computação) - Centro de Ciências e Tecnologia, Universidade Federal de Campina Grande, Campina Grande, 2002. Disponível em:

http://docs.computacao.ufcg.edu.br/posgraduacao/dissertacoes/2002/Dissertacao_BenitzDeSouzaVasconcelos.pdf. Acesso em: 10 ago. 2022.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**: practical machine learning tools and techniques. 3. ed. United States: Elsevier, 2011. ISSN 0453-4662.

APÊNDICE 1 - SIMPLIFICAÇÃO DE VALORES DO ATRIBUTO NOME JURÍDICO NO BOU

Original	DESCUMPRIR DECISAO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS DE URGENCIA PREVISTAS NESTA LEI
Simplificado	DESCUMBRIR ORDEM JUDICIAL DE MEDIDA PROTETIVA
Original	LESAO CORPORAL - SE A LESAO FOR PRATICADA CONTRA AUTORIDADE OU AGENTE DESCRITO NOS ARTS. 142 E 144
Simplificado	LESAO CORPORAL CONTRA AUTORIDADE OU AGENTE
Original	ESTA LEI CRIA MECANISMOS PARA COIBIR E PREVENIR A VIOLENCIA DOMESTICA E FAMILIAR CONTRA A MULHER, NOS TERMOS DO 8º DO ART. 226 DA CONSTITUICAO FEDERAL
Simplificado	VIOLENCIA CONTRA MULHER MARIA DA PENHA
Original	OFERECER, TROCAR, DISPONIBILIZAR, TRANSMITIR, VENDER OU EXPOR A VENDA, DISTRIBUIR, PUBLICAR OU DIVULGAR, POR QUALQUER MEIO - INCLUSIVE POR MEIO DE COMUNICACAO DE MASSA OU SISTEMA DE INFORMATICA OU TELEMATICA -, FOTOGRAFIA, VIDEO OU OUTRO REGISTRO AUDIOVISUAL QUE CONTENHA CENA DE ESTUPRO OU DE ESTUPRO DE VULNERAVEL OU QUE FAÇA APOLOGIA OU INDUZA A SUA PRATICA, OU, SEM O CONSENTIMENTO DA VITIMA, CEN
Simplificado	MANTER OU COMERCIALIZAR FOTOS E/OU VIDEOS DE ESTUPRO DE VULNERÁVEL
Original	AFASTAR OU PROCURA AFASTAR LICITANTE, POR MEIO DE VIOLENCIA, GRAVE AMEACA, FRAUDE OU OFERECIMENTO DE VANTAGEM DE QUALQUER TIPO
Simplificado	VIOLENCIA PSICOLOGICA
Original	FAVORECIMENTO DA PROSTITUICAO OU OUTRA FORMA DE EXPLORACAO SEXUAL COM VIOLENCIA / AMEACA
Simplificado	FAVORECER PROSTITUICAO
Original	ESTUPRO OU ATENTADO VIOLENTO AO PUDOR RESULTA LESAO CORPORAL GRAVE OU SE VITIMA MENOR COM MAIS DE 14
Simplificado	ESTUPRO/ATENTADO VIOLENTO AO PUDOR +14 LESAO GRAVE
Original	SE A INJRIA CONSISTE EM VIOLNCIA OU VIAS DE FATO, QUE, POR SUA NATUREZA OU MEIO EMPREGADO, SE CONSIDEREM AVILTANTES
Simplificado	INJURIA/VIAS DE FATO
Original	OFERECER, TROCAR, DISPONIBILIZAR, TRANSMITIR, DISTRIBUIR, PUBLICAR OU DIVULGAR POR QUALQUER MEIO, INCLUSIVE POR MEIO DE SISTEMA DE INFORMATICA OU TELEMATICO, FOTOGRAFIA, VIDEO OU OUTRO REGISTRO QUE CONTENHA CENA DE SEXO EXPLICITO OU PORNOGRAFICA ENVOLVENDO CRIANCA OU ADOLESCENTE
Simplificado	MANTER OU COMERCIALIZAR FOTOS/VIDEOS SEXO EXPLICITO COM MENOR
Original	CONDUZIR VEICULO AUTOMOTOR COM CAPACIDADE PSICOMOTORA ALTERADA EM RAZAO DA INFLUENCIA DE ALCOOL OU DE OUTRA SUBSTANCIA PSICOATIVA QUE DETERMINE DEPENDENCIA
Simplificado	CONDUZIR VEICULO SOB INFLUENCIA DE SUBSTANCIA PSICOATIVA
Original	SE DA PRATICA DO CRIME PREVISTO NO CAPUT RESULTAR LESAO CORPORAL DE NATUREZA GRAVE, E AS CIRCUNSTANCIAS DEMONSTRAREM QUE O AGENTE NAO QUIS O RESULTADO NEM ASSUMIU O RISCO DE PRODUZI-LO, A PENNA PRIVATIVA DE LIBERDADE E DE RECLUSAO, DE 3 (TRES) A 6 (SEIS) ANOS, SEM PREJUIZO DAS OUTRAS PENAS PREVISTAS NESTE ARTIGO
Simplificado	LESAO CORPORAL
Original	SE DA INVASAO RESULTAR A OBTENCAO DE CONTEUDO DE COMUNICACOES ELETRONICAS PRIVADAS, SEGREDOS COMERCIAIS OU INDUSTRIAIS, INFORMACOES SIGILOSAS, ASSIM DEFINIDAS EM LEI, OU O CONTROLE REMOTO NAO AUTORIZADO DO DISPOSITIVO INVADIDO

Simplificado	INVASAO NAO AUTORIZADA DE EQUIPAMENTO ELETRONICO
Original	ART. 91. RETER OU UTILIZAR CARTAO MAGNETICO, QUALQUER MEIO ELETRONICO OU DOCUMENTO DE PESSOA COM DEFICIENCIA DESTINADOS AO RECEBIMENTO DE BENEFICIOS, PROVENTOS, PENSOES OU REMUNERACAO OU A REALIZACAO DE OPERACOES FINANCEIRAS, COM O FIM DE OBTER VANTAGEM INDEVIDA PARA SI OU PARA OUTREM:
Simplificado	RETER OU UTILIZAR DOCUMENTOS OU CARTAO MAGNETICO DE RECEBIMENTO DE BENEFICIOS POR IDOSO
Original	PERSEGUICAO - ART. 147-A. PERSEGUIR ALGUÉM, REITERADAMENTE E POR QUALQUER MEIO, AMEACANDO-LHE A INTEGRIDADE FISICA OU PSICOLOGICA,
Simplificado	PERSEGUICAO COM AMEACA
Original	DAR PARTO ALHEIO COMO PROPRIO, REGISTRAR COMO SEU O FILHO DE OUTREM, OCULTAR RECEM-NASCIDO OU SUBSTITUI-LO, SUPRIMINDO OU ALTERANDO DIREITO INERENTE AO ESTADO CIVIL:
Simplificado	DAR PARTO ALHEIO COMO PRÓPRIO
Original	OFERECER DROGA, EVENTUALMENTE E SEM OBJETIVO DE LUCRO, A PESSOA DE SEU RELACIONAMENTO, PARA JUNTOS A
Simplificado	OFERECER DROGAS
Original	FAVORECIMENTO DA PROSTITUICAO OU DE OUTRA FORMA DE EXPLORACAO SEXUAL DE CRIANCA OU ADOLESCENTE OU DE VULNERAVEL
Simplificado	FAVORECIMENTO A PROSTITUICAO DE MENOR OU VULNERAVEL
Original	INVADIR OU ADENTRAR, CLANDESTINA OU ASTUCIOSAMENTE, OU A REVELIA DA VONTADE DO OCUPANTE, IMOVEL ALHEIO OU SUAS DEPENDENCIAS, OU NELE PERMANECER NAS MESMAS CONDICOES, SEM DETERMINACAO JUDICIAL OU FORA DAS CONDICOES ESTABELECIDAS EM LEI
Simplificado	INVASAO DE DOMICILIO
Original	FAVORECIMENTO DA PROSTITUICAO OU OUTRA FORMA DE EXPLORACAO SEXUAL SE O AGENTE TEM DEVER DE CUIDADO
Simplificado	FAVORECIMENTO A PROSTITUICAO DE PESSOA SOBRE CUIDADO
Original	ART. 89. APROPRIAR-SE DE OU DESVIAR BENS, PROVENTOS, PENSÃO, BENEFICIOS, REMUNERACAO OU QUALQUER OUTRO RENDIMENTO DE PESSOA COM DEFICIENCIA
Simplificado	APROPRIACAO INDEBITA CONTRA PESSOA COM DEFICIENCIA
Original	NA HIPOTHESE DO 30, AUMENTA-SE A PENA DE UM A DOIS TERCOS SE HOUVER DIVULGACAO, COMERCIALIZACAO OU TRANSMISSAO A TERCEIRO, A QUALQUER TITULO, DOS DADOS OU INFORMACOES OBTIDOS.
Simplificado	DISPONIBILIZAR DADOS NÃO AUTORIZADOS PARA TERCEIROS
Original	SE DA INVASAO RESULTAR A OBTENCAO DE CONTEUDO DE COMUNICACOES ELETRONICAS PRIVADAS, SEGREDOS COMERCIAIS OU INDUSTRIAIS, INFORMACOES SIGILOSAS, ASSIM DEFINIDAS EM LEI, OU O CONTROLE REMOTO NAO AUTORIZADO DO DISPOSITIVO INVADIDO
Simplificado	INVASAO NAO AUTORIZADA DE DISPOSITIVO ELETRONICO
Original	ART. 244-B. CORROMPER OU FACILITAR A CORRUPCAO DE MENOR DE 18 (DEZOITO) ANOS, COM ELE PRATICANDO INFRACAO PENAL OU INDUZINDO-O A PRATICA-LA
Simplificado	CORROMPER OU FACILITAR A CORRUPCAO DE MENORES
Original	II - DE METADE, SE O AGENTE E ASCENDENTE, PADRASTO OU MADRASTA, TIO, IRMAO, CONJUGE, COMPANHEIRO, TUTOR, CURADOR, PRECEPTOR OU EMPREGADOR DA VITIMA OU POR QUALQUER OUTRO TITULO TIVER AUTORIDADE SOBRE ELA
Simplificado	ABUSO SEXUAL DE VULNERAVEL

APÊNDICE 2 – REGRAS GERADAS PELO ALGORITMO CN2 COM A TÉCNICA DE SELEÇÃO DE ATRIBUTOS CFS

Nº	Se (If)	Então (Then)	Distribuição	Probabilidades	Qualidade
0	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=3 - 5	morte=S	[0, 2]	25 : 75	-0.00
1	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=7 - 9	morte=S	[0, 1]	33 : 67	-0.00
2	LESAO CORPORAL DE NATUREZA GRAVISSIMA=1 - 3	morte=S	[0, 8]	10 : 90	-0.00
3	CONSTRANGIMENTO ILEGAL=1 - 3	morte=S	[0, 1]	33 : 67	-0.00
4	DANO QUALIFICADO=1 - 3	morte=S	[0, 1]	33 : 67	-0.00
5	LESAO CORPORAL DE NATUREZA GRAVE=1 - 3	morte=S	[0, 4]	17 : 83	-0.00
6	VIAS DE FATO=1 - 3 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠< 1	morte=N	[1, 0]	67 : 33	-0.00
7	VIAS DE FATO=1 - 3 AND AMEACA≠1 - 3	morte=N	[8, 0]	90 : 10	-0.00
8	INJURIA=1 - 3 AND AMEACA=< 1 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠< 1	morte=N	[3, 0]	80 : 20	-0.00
9	INJURIA=1 - 3 AND AMEACA=< 1	morte=N	[9, 1]	83 : 17	-0.469
10	INJURIA=1 - 3 AND VIAS DE FATO=1 - 3	morte=N	[5, 1]	75 : 25	-0.650
11	INJURIA=1 - 3 AND AMEACA=1 - 3	morte=N	[16, 4]	77 : 23	-0.722
12	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠1 - 3 AND AMEACA=< 1	morte=S	[23, 69]	26 : 74	-0.811
13	VIAS DE FATO=1 - 3	morte=S	[2, 4]	38 : 62	-0.918
14	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠1 - 3 AND AMEACA≠3 - 5	morte=N	[30, 16]	65 : 35	-0.932
15	AMEACA=1 - 3	morte=N	[9, 6]	59 : 41	-0.971
16	AMEACA≠3 - 5	morte=N	[36, 24]	60 : 40	-0.971
17	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=1 - 3	morte=S	[1, 1]	50 : 50	-1.000
18	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=< 1	morte=S	[1, 1]	50 : 50	-1.000
19	TRUE	morte=N	[144, 144]	50 : 50	-1.000

APÊNDICE 3 – REGRAS GERADAS PELO ALGORITMO CN2 COM A TÉCNICA DE SELEÇÃO DE ATRIBUTOS INFO GAIN

Nº	Se (If)	Então (Then)	Distribuição	Probabilidades	Qualidade
0	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=3 - 5	morte=S	[0, 2]	25 : 75	-0.00
1	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=7 - 9	morte=S	[0, 1]	33 : 67	-0.00
2	LESAO CORPORAL DE NATUREZA GRAVISSIMA=1 - 3	morte=S	[0, 8]	10 : 90	-0.00
3	LESAO CORPORAL DE NATUREZA GRAVE=1 - 3	morte=S	[0, 4]	17 : 83	-0.00
4	DESCUMPRIR DECISÃO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS DE URGÊNCIA PREVISTAS NESTA LEI=1 - 3 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠< 1	morte=S	[0, 1]	33 : 67	-0.00
5	VIAS DE FATO=1 - 3 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠< 1	morte=N	[1, 0]	67 : 33	-0.00
6	VIAS DE FATO=1 - 3 AND AMEACA≠1 - 3	morte=N	[8, 0]	90 : 10	-0.00
7	DESCUMPRIR DECISÃO JUDICIAL QUE DEFERE MEDIDAS PROTETIVAS DE URGÊNCIA PREVISTAS NESTA LEI=1 - 3 AND AMEACA≠1 - 3	morte=S	[0, 2]	25 : 75	-0.00
8	AMEACA=3 - 5 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠1 - 3	morte=N	[1, 0]	67 : 33	-0.00
9	INJURIA=1 - 3 AND AMEACA=≠< 1 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠< 1	morte=N	[3, 0]	80 : 20	-0.00
10	INJURIA=1 - 3 AND boletins=1 - 3 AND VIAS DE FATO=1 - 3	morte=N	[4, 0]	83 : 17	-0.00
11	INJURIA=1 - 3 AND AMEACA=≠< 1	morte=N	[9, 1]	83 : 17	-0.469
12	INJURIA=1 - 3 AND boletins=1 - 3 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠1 - 3	morte=N	[9, 2]	77 : 23	-0.684
13	VIAS DE FATO=1 - 3 AND boletins=1 - 3	morte=S	[1, 4]	29 : 71	-0.722
14	VIAS DE FATO=1 - 3 AND INJURIA≠1 - 3	morte=N	[1, 0]	67 : 33	-0.00
15	INJURIA=1 - 3 AND AMEACA=1 - 3 AND LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=1 - 3 AND boletins≠1 - 3	morte=N	[1, 0]	67 : 33	-0.00
16	AMEACA=≠< 1 AND boletins≠1 - 3	morte=S	[0, 4]	17 : 83	-0.00
17	INJURIA=1 - 3 AND boletins=1 - 3	morte=N	[3, 1]	67 : 33	-0.811
18	AMEACA=1 - 3 AND INJURIA=1 - 3 AND VIAS DE FATO≠1 - 3	morte=N	[3, 1]	67 : 33	-0.811
19	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠1 - 3 AND AMEACA≠1 - 3	morte=S	[23, 68]	26 : 74	-0.816
20	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR≠1 - 3 AND boletins=1 - 3	morte=N	[28, 14]	66 : 34	-0.918
21	boletins=1 - 3 AND AMEACA≠< 1	morte=N	[6, 3]	64 : 36	-0.918
22	AMEACA=≠< 1	morte=N	[36, 22]	62 : 38	-0.958
23	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=1 - 3 AND AMEACA≠3 - 5	morte=N	[3, 2]	57 : 43	-0.971
24	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=1 - 3	morte=S	[1, 1]	50 : 50	-1.000
25	LESAO CORPORAL - VIOLENCIA DOMESTICA E FAMILIAR=≠< 1	morte=S	[3, 3]	50 : 50	-1.000
26	TRUE	morte=N	[144, 144]	50 : 50	-1.000