

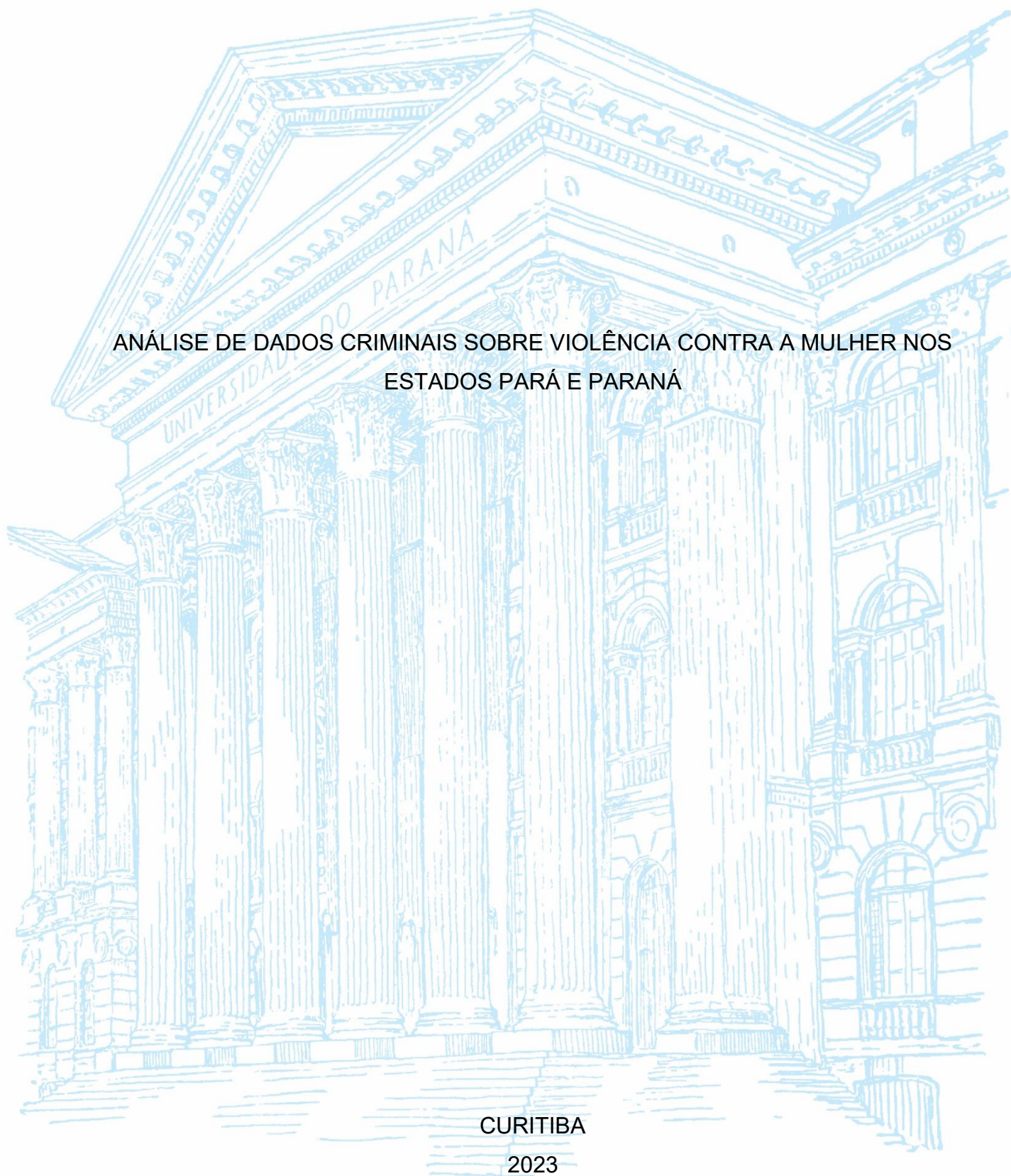
UNIVERSIDADE FEDERAL DO PARANÁ

MARIANA KNISS

ANÁLISE DE DADOS CRIMINAIS SOBRE VIOLÊNCIA CONTRA A MULHER NOS
ESTADOS PARÁ E PARANÁ

CURITIBA

2023



MARIANA KNISS

ANÁLISE DE DADOS CRIMINAIS SOBRE VIOLÊNCIA CONTRA A MULHER NOS
ESTADOS PARÁ E PARANÁ

Trabalho de Conclusão de Curso apresentado ao curso de Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Gestão da Informação.

Orientadora: Profa. Dra. Denise Fukumi Tsunoda

CURITIBA

2023

Dedico essa pesquisa a todas as meninas e mulheres que tiveram a liberdade condicionada ao desejo do outro. Em especial, àquelas que tiveram suas vidas e corpos invadidos, manipulados ou ceifados pela violência do patriarcado.

AGRADECIMENTOS

Primeiramente, agradeço à Universidade Federal do Paraná pelos anos proporcionados de aprendizado, conhecimento, relacionamentos que me permitiram a oportunidade de estar em contato com as mais diversas realidades e contextos sociais, que, tenho certeza, mudaram a maneira como enxergo e me relaciono com a sociedade.

Agradeço imensamente à minha orientadora Denise Fukumi Tsunoda, primeiramente pela dedicação em todas as disciplinas que tive oportunidade de ser sua aluna, as quais me fizeram encantar pelo mundo dos dados e me permitiram delinear meu futuro profissional. Ainda, agradeço por seus direcionamentos durante o desenvolvimento deste trabalho, pelo incentivo nos momentos em que me senti perdida, pelo conhecimento compartilhado, pela parceira sempre disponível.

À minha família, registro minha gratidão por acreditarem e incentivarem que eu voltasse a estudar, pela vibração a cada aprovação, seja no vestibular, seja nas matérias e, principalmente, por serem minha base emocional e pelos valores que me formaram como cidadã.

Aos amigos que fiz durante a trajetória no curso de Gestão da Informação, agradeço pelos momentos de descontração no hall do Sociais Aplicadas, pelo companheirismo que desenvolvemos durante os estágios, pela parceira de estudos e pelo apoio nos momentos em que o desânimo estudantil aparecia.

Gratidão às amigas e amigos que caminham junto desde antes de meu ingresso na UFPR, em especial aqueles que sempre demonstraram seu amor por essa Universidade e incutiram em mim a vontade de ser uma aluna federal, por todo incentivo e comemoração a cada conquista, pelas conversas que me fizeram ver o mundo de uma maneira diferente.

Por fim, quero demonstrar minha mais pura gratidão ao meu amor, Michael, por sua incansável parceria. Seu apoio durante minha graduação, e bem antes dela também, mudou a minha vida! Agradeço pelo incentivo aos estudos, por ser meu suporte nos momentos em que eu não acreditava em mim, por todos os esforços que fez para que eu chegasse até aqui.

“Não serei livre enquanto alguma mulher for prisioneira, mesmo que as
correntes dela sejam diferentes das minhas.”

(Audre Lorde)

RESUMO

Aborda a problemática da violência contra a mulher por meio de uma abordagem centrada na análise de dados, utilizando registros criminais das delegacias dos estados do Pará e Paraná no período de 2018 a 2022. Objetiva desenvolver análise crítica sobre os dados criminais de violência contra a mulher. Apresenta referencial teórico sobre gestão da informação, análise de dados, descoberta de conhecimento em bases de dados, violência contra a mulher, feminicídio e a relação entre análise de dados e violência contra a mulher. Caracteriza o estudo como exploratório, experimental e quantitativo. Define o processo KDD como base metodológica para o desenvolvimento da pesquisa, cuja execução se dá através das etapas pré-processamento, processamento e pós-processamento de dados. Para tanto, utiliza ferramentas como o Microsoft Excel, Python, Pyspark, Google Colab e Orange Data Mining. Aplica os algoritmos de aprendizado de máquina Random Forest e Multilayer Perceptron para a tarefa de classificação. Apresenta os resultados da estatística descritiva e da aplicação dos algoritmos de mineração de dados. Avalia os algoritmos de predição com base nas métricas estatísticas disponíveis na ferramenta Orange Data Mining. Discorre sobre a contribuição da pesquisa na elaboração de políticas públicas de enfrentamento à violência contra a mulher.

Palavras-chave: Análise de dados. Violência contra a mulher. Gestão da Informação.

KDD. Random Forest. Multilayer Perceptron.

ABSTRACT

It approaches the subject of violence against women from the perspective of data analysis. To this end, it analyzes data from criminal records from police stations in the states of Pará and Paraná between 2018 and 2022. It aims to develop a critical analysis of criminal data on violence against women. It presents a theoretical framework on information management, data analysis, knowledge discovery in databases, violence against women, femicide and the relationship between data analysis and violence against women. It characterizes the study as exploratory, experimental and quantitative. It considers the KDD process as the methodological basis for the development of statistical analysis and data mining. It defines the execution of the KDD process through the stages of data pre-processing, processing and post-processing. It uses tools such as Microsoft Excel, Python, Pyspark and Google Colab and Orange Data Mining. It applies the Random Forest and Multilayer Perceptron machine learning algorithms to the classification task. It presents the results of descriptive statistics and the application of data mining algorithms. It evaluates the prediction algorithms based on the statistical metrics available in the Orange Data Mining tool. It discusses the contribution of the research to the development of public policies to combat gender violence.

Keywords: Data analysis. Violence against woman. Information Management. KDD. Random Forest. Multilayer Perceptron.

LISTA DE FIGURAS

FIGURA 1 – PROCESSO DE GERENCIAMENTO DA INFORMAÇÃO	25
FIGURA 2 – CICLO PROCESSUAL DE GESTÃO DA INFORMAÇÃO	26
FIGURA 3 – CICLO DE VIDA DA INFORMAÇÃO	26
FIGURA 4 – PROCESSO KDD	32
FIGURA 5 – TAREFAS DE MINERAÇÃO DE DADOS	34
FIGURA 6 – INSTALAÇÃO DE BIBLIOTECAS, MÓDULOS E PACOTES	47
FIGURA 7 – PROCESSO DE PREPARAÇÃO DA BASE DE DADOS.....	51
FIGURA 8 – INSTALAÇÃO DO PYSPARK E CONFIGURAÇÃO DO AMBIENTE	51
FIGURA 9 – WIDGETS DO MÓDULO 'DATA' DO ORANGE DATA MINING	60
FIGURA 10 – WIDGETS DO MÓDULO 'TRANSFORM' DO ORANGE DATA MINING	61
FIGURA 11 – TELA DE CONFIGURAÇÃO DO WIDGET 'CREATE CLASS'	62
FIGURA 12 – TELA DE CONFIGURAÇÃO DO WIDGET 'PREPROCESS'	63
FIGURA 13 – TELA DE CONFIGURAÇÃO DO WIDGET 'SELECT COLUMNS'	64
FIGURA 14 – TELA DE CONFIGURAÇÃO DO WIDGET 'DATA SAMPLER'	65
FIGURA 15 – WIDGETS DO MÓDULO 'MODEL' DO ORANGE DATA MINING.....	66
FIGURA 16 – TELA DE CONFIGURAÇÃO DO WIDGET 'RANDOM FOREST'	67
FIGURA 17 – TELA DE CONFIGURAÇÃO DO WIDGET 'NEURAL NETWORK'	68
FIGURA 18 – WIDGETS DO MÓDULO 'EVALUATE' DO ORANGE DATA MINING	69
FIGURA 19 – TELA DE CONFIGURAÇÃO DO WIDGET 'TEST AND SCORE'	70
FIGURA 20 – WORKFLOW DESENVOLVIDO NO ORANGE DATA MINING	71
FIGURA 21 – MATRIZ DE CONFUSÃO DO ALGORITMO MLP	91
FIGURA 21 – MATRIZ DE CONFUSÃO DO ALGORITMO RF	92

LISTA DE GRÁFICOS

GRÁFICO 1 – EVOLUÇÃO DOS NÍVEIS DE VITIMIZAÇÃO ENTRE MULHERES NO BRASIL. SÉRIE HISTÓRICA, 2017-2023.	18
GRÁFICO 2 – TOTAL DE VÍTIMAS DE FEMINICÍDIO NO 1º SEMESTRE DE CADA ANO NO BRASIL, 2019 A 2023	19
GRÁFICO 3 – CONTAGEM DE PUBLICAÇÕES POR ANO DA BASE SCOPUS	22
GRÁFICO 4 – REGISTROS DE OCORRÊNCIAS POR ANO	73
GRÁFICO 5 – REGISTROS DE OCORRÊNCIAS POR MÊS	74
GRÁFICO 6 – REGISTROS DE OCORRÊNCIAS POR SEMANA.....	74
GRÁFICO 7 – REGISTROS DE OCORRÊNCIAS POR PERÍODO DO DIA	75
GRÁFICO 8 – DISPERSÃO DE IDADES DAS VÍTIMAS AGRUPADAS POR ANO .	78
GRÁFICO 9 – REGISTROS DE OCORRÊNCIAS POR COR DA PELE DA VÍTIMA	78
GRÁFICO 10 – REGISTROS DE OCORRÊNCIAS POR PROFISSÃO DA VÍTIMA .	80
GRÁFICO 11 – DISPERSÃO DE IDADES DOS AUTORES AGRUPADAS POR ANO	81

LISTA DE QUADROS

QUADRO 1 – TIPOS DOS DOCUMENTOS RECUPERADOS DA BASE LENS	21
QUADRO 2 – DISTINÇÃO ENTRE DADO, INFORMAÇÃO E CONHECIMENTO	24
QUADRO 3 – ATRIBUTOS DA BASE DE VIOLÊNCIA CONTRA A MULHER	49
QUADRO 4 – REGISTROS DE OCORRÊNCIAS POR GRAU DE INSTRUÇÃO DA VÍTIMA	80
QUADRO 5 – REGISTROS DE OCORRÊNCIAS POR COR DA PELE DO AUTOR	82
QUADRO 6 – REGISTROS DE OCORRÊNCIAS POR GRAU DE INSTRUÇÃO DO AUTOR	82
QUADRO 7 – REGISTROS DE OCORRÊNCIAS POR PROFISSÃO DO AUTOR...	83
QUADRO 8 – REGISTROS DE OCORRÊNCIAS POR GRAU DE RELACIONAMENTO COM A VÍTIMA	84
QUADRO 9 – REGISTROS DE OCORRÊNCIAS POR MEIO EMPREGADO	84
QUADRO 10 – REGISTROS DE OCORRÊNCIAS POR LOCAL DE OCORRÊNCIA	85
QUADRO 11 – REGISTROS DE OCORRÊNCIAS POR REGISTRO.....	86

LISTA DE TABELAS

TABELA 1 – REGISTROS DE OCORRÊNCIAS POR FAIXA HORÁRIA.....	75
TABELA 2 – MEDIDAS DE RESUMO DO ATRIBUTO IDADE POR ANO	76
TABELA 3 – CLASSIFICAÇÃO E REGISTROS DO ATRIBUTO FAIXA ETÁRIA DA VÍTIMA	77
TABELA 4 – RESULTADOS DA EXECUÇÃO DO ALGORITMO MLP NO MÉTODO DE AMOSTRAGEM <i>BOOTSTRAP</i>	88
TABELA 5 – RESULTADOS DA EXECUÇÃO DO ALGORITMO MLP NO MÉTODO DE AMOSTRAGEM <i>CROSS VALIDATION</i>	89
TABELA 6 – RESULTADOS DA EXECUÇÃO DO ALGORITMO RF NO MÉTODO DE AMOSTRAGEM <i>BOOTSTRAP</i>	89
TABELA 7 – RESULTADOS DA EXECUÇÃO DO ALGORITMO RF NO MÉTODO DE AMOSTRAGEM <i>CROSS VALIDATION</i>	90
TABELA 8 – RESULTADOS DA EXECUÇÃO DOS ALGORITMOS RF E MLP	91

LISTA DE ABREVIATURAS OU SIGLAS

CID-10	- Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde
DM	- Data Mining
FBSP	- Fórum Brasileiro de Segurança Pública
GC	- Gestão do Conhecimento
GI	- Gestão da Informação
KDD	- Knowledge Discovery in Databases
MLP	- Multilayer Perceptron
PNAINFO	- Política Nacional de Dados e Informações relacionadas à Violência contra as Mulheres
RF	- Random Forest
SIM	- Sistema de Informações sobre Mortalidade
SUS	- Sistema Único de Saúde
UFPR	- Universidade Federal do Paraná
UNODC	- Escritório das Nações Unidas sobre Drogas e Crime
UN WOMEN	- Entidade das Nações Unidas para a Igualdade de Gênero e o Empoderamento das Mulheres
VCM	- Violência Contra a Mulher
VD	- Violência Doméstica

SUMÁRIO

1 INTRODUÇÃO	14
1.1 DEFINIÇÃO DO PROBLEMA.....	15
1.2 OBJETIVOS	17
1.2.1 Objetivo geral	17
1.2.2 Objetivos específicos.....	17
1.3 JUSTIFICATIVA	18
1.4 DELIMITAÇÃO DA PESQUISA.....	23
1.5 ESTRUTURA DO DOCUMENTO.....	23
2 REVISÃO DE LITERATURA	24
2.1 GESTÃO DA INFORMAÇÃO	24
2.2 ANÁLISE DE DADOS.....	28
2.3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (KDD)	29
2.3.1 Mineração de dados	32
2.3.1.1 Tarefas de mineração de dados	33
2.4 VIOLÊNCIA CONTRA A MULHER.....	35
2.4.1 Femicídio	38
2.5 RELAÇÕES ENTRE ANÁLISE DE DADOS E VIOLÊNCIA CONTRA A MULHER	
41	
3 ENCAMINHAMENTOS METODOLÓGICOS	45
3.1 CARACTERIZAÇÃO DA PESQUISA	45
3.2 MATERIAIS E MÉTODOS.....	46
3.2.1 Base de dados	48
3.2.2 Pré-processamento	50
3.2.3 Processamento	58
3.2.4 Pós-processamento	68
4 RESULTADOS E DISCUSSÃO	72
4.1 ESTATÍSTICA DESCRITIVA.....	72
4.2 MINERAÇÃO DE DADOS	87
4.3 ANÁLISE CRÍTICA SOBRE A ANÁLISE DE DADOS CRIMINAIS DE VIOLÊNCIA CONTRA A MULHER.....	92
5 CONSIDERAÇÕES FINAIS	96
REFERÊNCIAS.....	99

1 INTRODUÇÃO

O massivo acesso à informação experienciado atualmente serve de pano de fundo para maior visibilidade a um problema de saúde pública que interfere em todas as esferas sociais: a violência contra a mulher. Nos últimos anos, estudiosos, governos e organizações internacionais têm despertado atenção ao tema, devido à sua magnitude e consequências sociais. Coletivos feministas, campanhas de conscientização e mídias sociais fomentam o campo informacional acerca dos crimes de ódio praticados contra meninas e mulheres.

Nesse cenário, a quantidade de dados referentes à violência contra a mulher também cresce a cada dia. Segundo Tsunoda e Constâncio (2023), embora o aumento do volume de dados ofereça oportunidades para compreender os fenômenos sociais e utilizar as informações descobertas na tomada de decisões, é incontestável que a quantidade, variedade e complexidade avassaladora desses dados constituem, por si só, um desafio para os pesquisadores, que precisam encontrar meios para realizar análises e identificar padrões de interesse.

A violência de gênero contra mulheres está enraizada na cultura da sociedade patriarcal, que possui estruturas desiguais de poder, e pode assumir diversas formas, como psicológica, física, moral e patrimonial (Tolosa; Chagas; Lima, 2020). De acordo com o relatório “*Violence against women prevalence estimates, 2018*”, estima-se que 30% da população feminina mundial acima de 15 anos tenha sido submetida à violência física ou sexual pelo parceiro íntimo ou sofrido violência sexual de um não parceiro - estranhos, conhecidos, amigos, colegas, professores, vizinhos, membros da família - havendo casos em que ocorreram ambas categorias de violência (WHO, 2021).

Embora a grande maioria das vítimas de homicídio registradas globalmente sejam do gênero masculino, as mulheres e meninas são, em todas as regiões do mundo, desproporcionalmente afetadas por homicídios na esfera privada, visto que mais da metade dos assassinatos do gênero feminino são praticados por parceiro íntimo ou familiar (UNODC, 2022).

O Brasil, que em 2015 ocupava o quinto lugar dentre 83 países com dados homogêneos sobre homicídios femininos (Waiselfisz, 2015), decretou nos últimos vinte anos duas importantes leis como forma de oposição e enfrentamento à violência contra a mulher: a Lei n.º 11.340/2006, conhecida como Lei Maria da Penha, e a Lei

n.º 13.104/2015, que tipifica o homicídio feminino no contexto da misoginia como feminicídio.

O feminicídio é o ápice de um *continuum* de diversas agressões praticadas contra a mulher, que, ao longo do tempo, vão sendo naturalizadas na sociedade, e é fortemente associado a situações de desigualdade e discriminação de gênero, privação econômica e masculinidade agressiva e machista (Roichman, 2020; Meneghel; Hirakata, 2011).

Conforme apontam Romeiro e Bezerra (2020), se a naturalização da violência contra a mulher atua como fator de invisibilidade do sofrimento feminino, embasar a discussão através de dados e informações tem o potencial de dar visibilidade ao tema e estimular a denúncia de agressões, bem como possibilitar a criação de políticas públicas efetivas que auxiliem no combate eficaz ao fenômeno, a fim de garantir direitos relacionados à proteção física e psicológica das mulheres.

Nesse contexto, faz-se atual e necessária a produção de pesquisas sobre a violência contra a mulher sob a ótica da análise de dados, para que seja possível obter um panorama claro e abrangente sobre a incidência e características desse tipo de crime. Tais estudos são capazes de fornecer informações valiosas para as autoridades legisladoras através da identificação de padrões e tendências sobre o fenômeno, possibilitando assim novos conhecimentos que sejam capazes de embasar a tomada de decisão quanto às políticas públicas de proteção ao direito e à vida das mulheres.

Desse modo, o presente trabalho objetiva a realização de uma análise crítica sobre a violência contra a mulher no Brasil, considerando dados criminais das Secretarias Públicas dos estados Pará e Paraná, a fim de identificar as limitações e desafios da análise de dados sobre a violência contra a mulher, fatores esses que impactam na geração de informações e descoberta de conhecimentos sobre o tema.

1.1 DEFINIÇÃO DO PROBLEMA

Ao analisar o impacto das informações geradas através dos sistemas de informação e comunicação pública, Campos e Almeida (2017) afirmam que é preciso considerar fundamentalmente o peso do caráter cultural e político que o fenômeno da violência assume na sociedade atual e a forma como influencia no processo de produção da informação. Desse modo, conforme os autores, os sistemas de

informação se tornam um importante mecanismo para identificar e diagnosticar incidências relativas a este fenômeno, pois, ao possibilitarem o mapeamento de casos através da construção de indicadores para dimensionamento do problema, influenciam o planejamento de iniciativas mais eficazes relacionadas ao seu enfrentamento.

Sendo a violência contra a mulher um evento multicausal, frequente e complexo (Caicedo-Roa; Cordeiro, 2023), a compreensão completa desse fenômeno e a implementação de medidas efetivas de prevenção exigem uma análise abrangente dos dados que o descrevem. Entretanto, através do levantamento bibliográfico realizado para o embasamento teórico desta pesquisa, observou-se que existem dificuldades enfrentadas ao analisar dados sobre o feminicídio.

Estimar o número global de assassinatos de mulheres e meninas relacionados ao gênero é desafiador (UNODC; UN Women, 2022). Um fator de impacto na problemática envolvendo dados sobre feminicídio se dá pela subnotificação dos casos. Pode-se apontar como uma das causas dessa subnotificação o fato de que, em alguns países, o feminicídio não é devidamente tipificado como crime, o que resulta em uma falta de registros oficiais e, conseqüentemente, na subestimação do problema. A subnotificação dos casos compromete a precisão dos dados disponíveis, dificultando a análise acurada do fenômeno e a identificação de padrões e tendências. Por outro lado, em países que preveem o feminicídio judicialmente, o problema dos pesquisadores se inicia na escolha da fonte de dados a ser analisada: bases provenientes dos sistemas de saúde ou da polícia. Cada fonte de dados apresenta vantagens na análise sob determinadas óticas, em detrimento de outras.

No Brasil, com base na Lei n.º 13.104, desde 2015 há a tipificação do feminicídio no código penal. Ainda assim, para que seja tipificado como feminicídio, o homicídio deve ser caracterizado pela misoginia, sendo muitas vezes identificado através de denúncias prévias de violência doméstica. Contudo, frequentemente, as vítimas não denunciam os crimes devido ao medo, à intimidação e à falta de confiança nas instituições responsáveis pela aplicação da lei, gerando subnotificação nos casos.

Outro fator de relevância na análise de dados sobre a violência contra a mulher é a falta de padronização na coleta e na categorização dos dados. Em muitos casos, os registros não incluem informações detalhadas sobre as circunstâncias do crime, as características das vítimas e dos agressores, e os fatores de risco associados. A ausência de uma abordagem unificada para a coleta e a categorização

de dados dificulta a comparação entre diferentes localidades e a identificação de fatores comuns que contribuem para o fenômeno.

Adicionalmente, outro fator a ser considerado ao analisar dados relativos ao feminicídio se refere à privacidade e à ética. Dados sobre vítimas e autores de crimes geralmente envolvem informações pessoais sensíveis, o que requer uma abordagem cuidadosa para garantir a confidencialidade e a proteção da identidade das vítimas. Esses pontos devem ser considerados ao conduzir a análise de dados sobre feminicídio, a fim de garantir a integridade e a segurança dos envolvidos.

Nesse contexto, onde é preciso identificar e analisar criticamente os fatores que comprometem a integridade, confiabilidade e disponibilidade de dados sobre o fenômeno, este trabalho visa responder à seguinte pergunta de pesquisa: **quais são as limitações e desafios da análise de dados sobre violência contra a mulher e como podem ser superados?**

1.2 OBJETIVOS

Como forma de nortear o desenvolvimento desta pesquisa é preciso definir objetivos que contribuam para a construção da solução do problema de pesquisa identificado. Para tanto, são definidos o objetivo geral e os específicos, conforme seções abaixo.

1.2.1 Objetivo geral

O principal objetivo desta pesquisa é realizar uma análise crítica sobre a análise de dados criminais de violência contra a mulher, através de estudo sobre ocorrências registradas nos estados Pará e Paraná, a fim de identificar as limitações e desafios da análise de dados sobre esse fenômeno.

1.2.2 Objetivos específicos

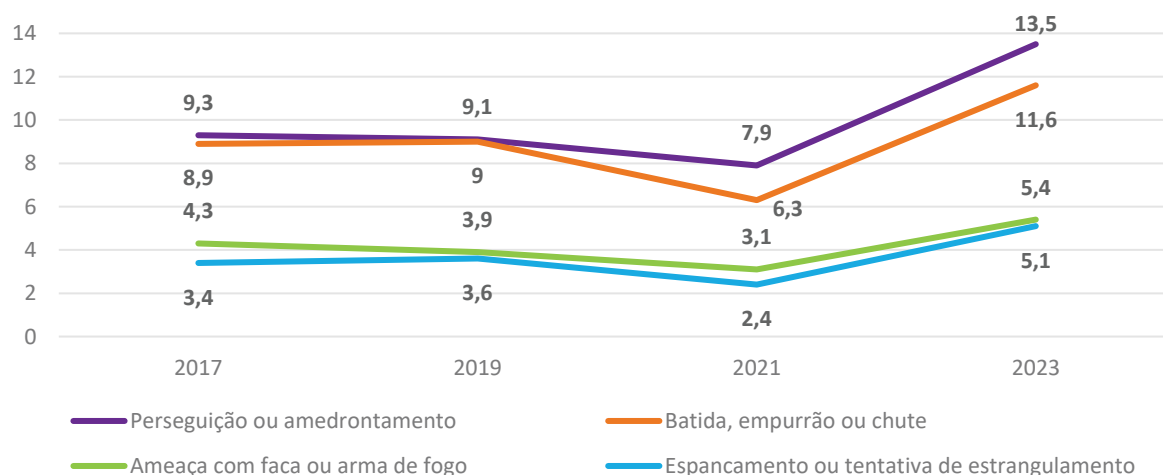
Para que seja possível atingir o objetivo geral deste trabalho, dele se derivam os objetivos específicos elencados abaixo:

- a) construir uma base unificada que reúna metadados de duas bases de dados sobre a temática em pauta;
- b) definir a ferramenta de mineração de dados, estabelecer critérios de análise, utilizar e comparar os métodos;
- c) avaliar resultados obtidos por meio de critérios estatísticos;
- d) alinhar os resultados obtidos com o levantamento de literatura e produzir *insights* de aplicação.

1.3 JUSTIFICATIVA

Segundo a quarta edição da pesquisa “Visível e Invisível: a Vitimização de Mulheres no Brasil”, publicada pelo Fórum Brasileiro de Segurança Pública (FBSP) em março de 2023, houve aumento acentuado em todas as formas de violência contra a mulher desde a primeira edição da pesquisa, como aponta o gráfico 1. Os resultados mostram que 33,6% das mulheres entrevistadas já sofreram violência física e/ou sexual por parte do parceiro, ou ex-parceiro, íntimo. O panorama apresentado pela pesquisa de vitimização é corroborado pelo 17º Anuário Brasileiro de Segurança Pública, publicado em julho de 2023, que compilou registros administrativos de todas as unidades federativas do país e constatou aumento de 2,9% nas taxas de agressões em contexto de violência doméstica, 6,1% nas taxas de feminicídio e 49,7% de registros de assédio sexual (FBSP, 2023).

GRÁFICO 1 – EVOLUÇÃO DOS NÍVEIS DE VITIMIZAÇÃO ENTRE MULHERES NO BRASIL. SÉRIE HISTÓRICA, 2017-2023.

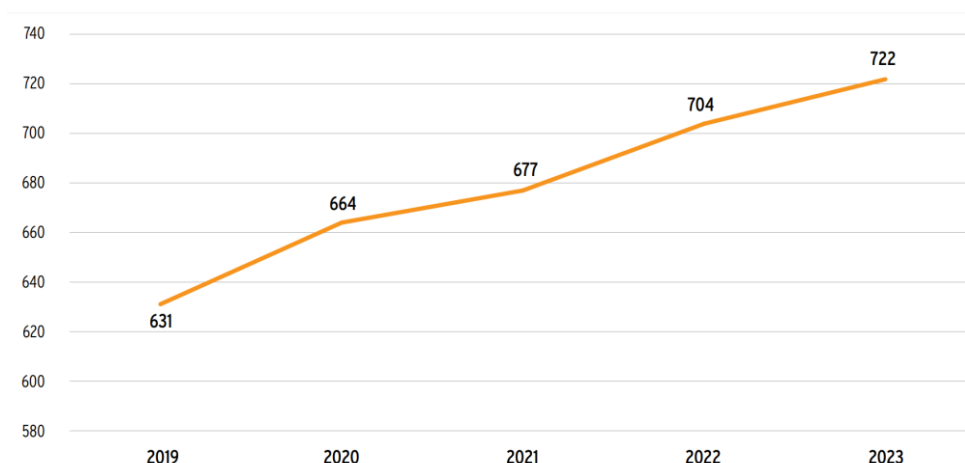


FONTE: Fórum Brasileiro de Segurança Pública (2023).

Segundo o relatório “*Gender-related killings of women and girls (femicide/feminicide)*”, elaborado pelo Escritório das Nações Unidas sobre Drogas e Crime (UNODC) e pela Entidade das Nações Unidas para a Igualdade de Gênero e o Empoderamento das Mulheres (UN Women), estima-se que, em 2021, 45.000 mulheres e meninas tenham sido mortas por parceiros íntimos ou por membros da família; esse número equivale a 56% do número mundial de homicídios femininos. O relatório ressalta que, embora os números sejam alarmantemente altos, a escala real do feminicídio deve ser muito maior, devido a inconsistências nas definições e critérios sobre o fenômeno, visto que, dos dados fornecidos pelos países signatários, 40 % não tinham informação suficiente para identificá-los como feminicídio, especialmente sobre os acontecidos na esfera pública (UNODC; UN Women, 2022).

De acordo com o relatório “Violência contra meninas e mulheres no 1º semestre de 2023”, também produzido pelo FBSP, que teve como fonte os boletins de ocorrência classificados com a qualificadora feminicídio pelas Polícias Civis dos Estados e Distrito Federal, os dados evidenciam que há crescimento contínuo das mortes de mulheres em razão do gênero feminino desde 2019 no Brasil, conforme o gráfico 2.

GRÁFICO 2 – TOTAL DE VÍTIMAS DE FEMINICÍDIO NO 1º SEMESTRE DE CADA ANO NO BRASIL, 2019 A 2023



FONTE: Fórum Brasileiro de Segurança Pública (2022).

Conforme Saracevic¹ (1974 apud Romeiro; Bezerra, 2020, p. 2), a “informação como serviço público pode servir como um instrumento para traçar diretrizes para a política e o planejamento”. Desse modo, a definição do tema da presente pesquisa se deu pela observação da urgência da sociedade em entender mais profundamente as variáveis que envolvem a violência contra a mulher, pois a luta pelos direitos das mulheres deve ser central na luta pela justiça social. Nesse sentido, desenvolver estudos que buscam encontrar padrões em dados sobre violência de gênero contra mulheres exerce relevância na vida de meninas e mulheres, bem como no desenvolvimento da ciência e no avanço em direção à igualdade social.

Para Caicedo-Roa, Bandeira e Cordeiro (2022, p.12), a violência de gênero é um fenômeno “complexo, multicausal e varia segundo o contexto histórico, cultural e social”. Assim sendo, é de extrema relevância a produção de pesquisa quanto ao tema sob a perspectiva da Gestão da Informação (GI), devido essa ser um campo de estudo multidisciplinar das Ciências Sociais. Para tanto, é possível se valer de conhecimentos adquiridos em disciplinas estudadas durante o curso como Métodos Quantitativos de Análise, Metodologia de Pesquisa, Introdução à Teoria da Informação, Mineração de Dados, Inteligência Artificial para Gestão da Informação, Produtos e Serviços de Informação, Competência Informacional e Análise da Informação e do Conhecimento.

Visando entender o panorama da produção científica acerca do tema, utilizou-se como estratégia de busca os descritores “análise de dados” e “violência contra a mulher” no acervo digital da Universidade Federal do Paraná (UFPR); a busca foi realizada no mês de agosto de 2023 e retornou 29 resultados. Todos os estudos retornados foram analisados individualmente, a fim de encontrar relação entre a presente pesquisa e outras já publicadas. Através da análise de cada um dos estudos, pode-se verificar que apenas dois deles possuíam aderência total ao tema, sendo um deles a dissertação intitulada “Proposta de fluxo de análise de dados preditiva em violência doméstica contra a mulher”, desenvolvida pelo aluno do Programa de Pós-Graduação em Gestão da Informação da UFPR, Christian Aristóteles da Silva Costa.

Em um segundo momento, visando maior especificidade no tema da pesquisa, realizou-se revisão bibliográfica nas bases de dados Lens e Scopus, através do acesso federado pelo portal de periódicos da CAPES, para identificar pesquisas

¹ SARACEVIC, T. Tecnologia da informação, sistemas de informação e informação como utilidade pública. **Ciência da Informação**, Rio de Janeiro, v. 3, n. 1, p. 57-67, 1974.

relacionadas ao feminicídio. Em ambas as bases, a busca, que ocorreu em agosto de 2023, foi feita pelos descritores “*data analysis*” e “*femicide*”; a partir dos documentos recuperados, foi feita uma análise bibliométrica para maior aprofundamento nos estudos já publicados.

Da base Lens, foram recuperados 97 documentos, com período de publicação entre 2003 e 2023. Dos 5 autores mais ativos, 2 deles são brasileiros e foram citados nesta pesquisa. O quadro 1 mostra a distribuição dos tipos de documentos recuperados através da busca.

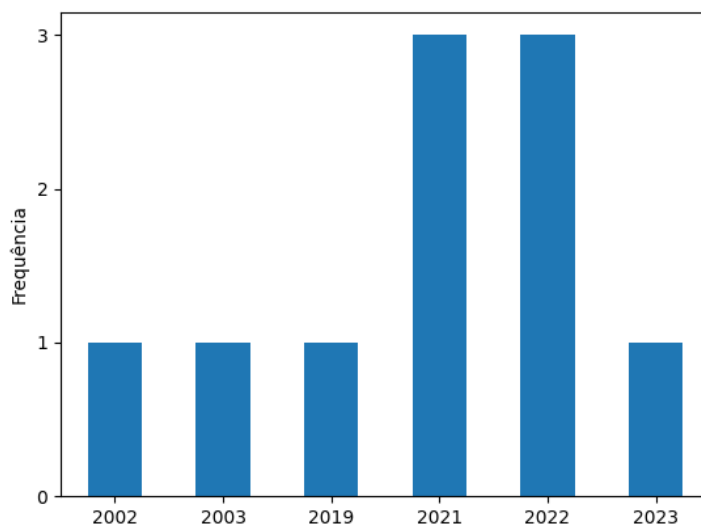
QUADRO 1 – TIPOS DOS DOCUMENTOS RECUPERADOS DA BASE LENS

Tipo do documento	Quantidade
Artigo de periódico	62
Dissertação	26
Livro	4
Relatório	1
Capítulo de livro	1
Não identificado	3

FONTE: A autora (2023).

Da base de dados Scopus, recuperou-se 10 documentos, sendo todos esses artigos de periódico e a maioria não tendo aderência total ao tema deste trabalho, por se tratarem de análises qualitativas de narrativas sobre feminicídio. A partir da análise bibliométrica dos 10 artigos recuperados, pode-se perceber um hiato na produção de estudos entre 2003 e 2019, conforme explicita o gráfico 3, e crescimento dos números a partir de 2021, evidenciando maior interesse em pesquisas sobre o tema.

GRÁFICO 3 – CONTAGEM DE PUBLICAÇÕES POR ANO DA BASE SCOPUS



FONTE: A autora (2023).

Baseando-se nas buscas ao acervo digital da UFPR e nas bases Lens e Scopus, cabe dizer que um motivo para a baixa quantidade de produção científica sobre o tema pode ser a dificuldade de acesso a dados sobre violência contra a mulher, visto a criticidade dos dados e informações que permeiam o fenômeno. Nesse sentido, tendo a autora dados disponíveis sobre o tema, a presente pesquisa ganha forte motivação, pois estudar os dados disponibilizados pelas Secretarias de Segurança Pública dos estados do Pará e Paraná dá sentido à parceria firmada com os órgãos, além de poder retornar às Secretarias e aos estados conhecimentos relevantes sobre os dados, que podem ser utilizados para a melhora das políticas públicas locais.

A partir do apresentado nesta seção, fica evidente a justificativa social para o desenvolvimento desta pesquisa, na medida em que pode servir de embasamento para melhor compreensão sobre a violência contra a mulher e os fatores que influenciam a análise de dados sobre o tema, e, conseqüentemente, contribuir com o desenvolvimento de políticas públicas de enfrentamento à violência contra a mulher que traduzam a realidade do fenômeno no país. De modo complementar, o presente trabalho conta ainda com a motivação acadêmica, visto o baixo volume de estudos publicados com aderência total a esse tema tão relevante para a sociedade, pois pode influenciar o desenvolvimento de novas pesquisas relacionados com a área de estudo.

1.4 DELIMITAÇÃO DA PESQUISA

A presente pesquisa visa explorar o tema análise de dados sobre violência contra a mulher, a fim de realizar uma análise crítica sobre as limitações e desafios enfrentados ao pesquisar o tema. Para tanto, delimitou-se a analisar registros criminais fornecidos pelas Secretarias de Segurança Pública dos estados Pará e Paraná, contendo ocorrências entre os anos 2018 a 2022.

1.5 ESTRUTURA DO DOCUMENTO

Este documento está dividido em cinco seções principais. A primeira seção contém a introdução, a definição do problema e questão de pesquisa, além dos objetivos gerais e específicos, justificativa e a delimitação da pesquisa.

Na segunda seção é apresentada a revisão de literatura a fim de conceituar os assuntos abordados na pesquisa, sendo eles: gestão da informação, análise de dados, descoberta de conhecimento em bases de dados, violência contra a mulher; além de, complementarmente, evidenciar a relação entre a análise de dados e violência contra a mulher.

Na seção três discorre-se sobre os encaminhamentos metodológicos percorridos no desenvolvimento deste estudo, tais como a caracterização metodológica da pesquisa, a apresentação dos materiais necessários e métodos aplicados sobre eles.

A quarta seção apresenta os resultados obtidos a partir do desenvolvimento da pesquisa e breve discussão sobre o tema.

Para finalizar, a quinta seção contém as considerações finais, onde é feita uma revisão quanto ao atingimento dos objetivos propostos, bem como a avaliação da contribuição da pesquisa e recomendação para trabalhos futuros.

2 REVISÃO DE LITERATURA

Esta seção se dedica a elucidar conceitos relevantes para a compreensão dos principais assuntos que compõem a problemática do presente trabalho. Para tanto, é apresentado referencial teórico quanto aos temas gestão da informação, análise de dados, descoberta de conhecimento em bases de dados, violência contra a mulher e feminicídio, além de uma seção que visa evidenciar a relação entre a análise de dados e violência contra a mulher, através da apresentação de estudos relativos ao tema.

2.1 GESTÃO DA INFORMAÇÃO

Em seu livro intitulado “Ecologia da Informação: por que só a tecnologia não basta para o sucesso na era da informação”, no qual discorre sobre a necessidade de uma abordagem holística para a gestão da informação, Davenport (1998, p. 173) define o gerenciamento informacional como um processo, “um conjunto estruturado de atividades que incluem o modo como as empresas obtêm, distribuem e usam a informação e o conhecimento”.

A fim de esclarecer o conceito do principal objeto do estudo, Davenport (1998) apresenta a distinção entre os três termos frequentemente usados para definir informação, a saber: dado, informação e conhecimento. Conforme o autor, é difícil, na prática, distinguir os três termos, pois estão fortemente relacionados e um serve de conexão para a compreensão do outro, conforme explicitado no quadro 2.

QUADRO 2 – DISTINÇÃO ENTRE DADO, INFORMAÇÃO E CONHECIMENTO

Dado	Informação	Conhecimento
Simples observações sobre o estado do mundo • Facilmente estruturado • 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 • Inclui reflexão, síntese, contexto • De difícil estruturação • De difícil captura em máquinas • Frequentemente tácito • De difícil transferência

FONTE: Davenport (1998, p. 18).

Para Davenport (1998), enxergar o gerenciamento informacional como um processo gera a possibilidade de mensuração e aperfeiçoamento do mesmo. Nesse sentido, o autor propõe o processo de gerenciamento da informação através de um modelo genérico que contém quatro etapas, identificadas na figura 1. A primeira etapa trata da identificação das necessidades informacionais; a segunda se preocupa com a obtenção da informação, através da seleção das fontes, além da exploração, classificação, formatação e estruturação das informações; a terceira versa sobre as estratégias de distribuição e comunicação da informação; por fim, a quarta etapa discorre sobre o uso da informação.

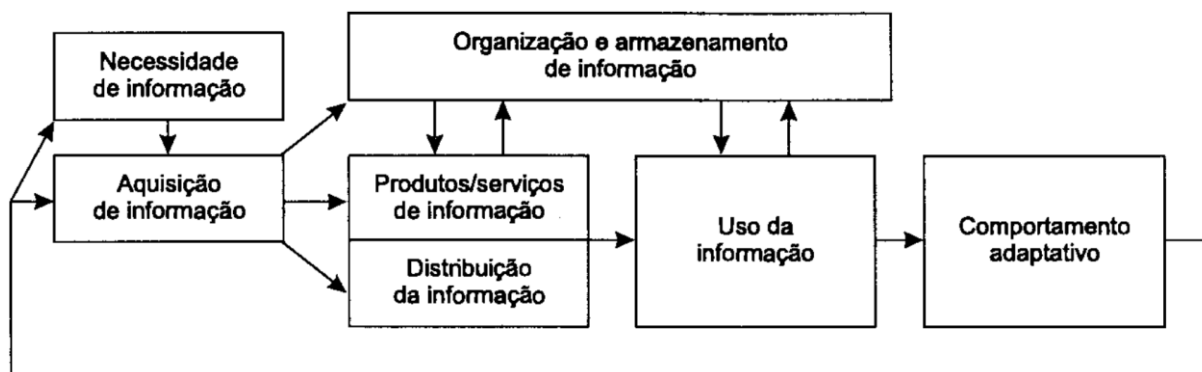
FIGURA 1 – PROCESSO DE GERENCIAMENTO DA INFORMAÇÃO



FONTE: Davenport (1998, p. 175).

Do mesmo modo, Choo (2003, p. 403) analisou o uso da informação organizacional em termos de necessidades, busca e uso informacionais e abordou a gestão da informação sob a perspectiva de “uma rede de processos que adquirem, criam, organizam, distribuem e usam a informação”. Para o autor, a GI deve ser analisada a partir de um ciclo contínuo de seis processos correlatos, evidenciado na figura 2, e relaciona o uso do ciclo processual de gestão da informação com a capacidade de aprendizado organizacional e a geração de conhecimento, para os indivíduos e a organização.

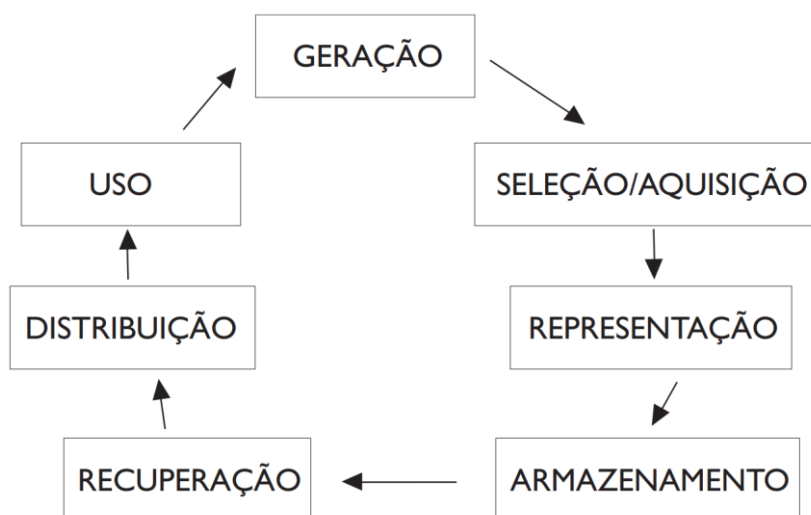
FIGURA 2 – CICLO PROCESSUAL DE GESTÃO DA INFORMAÇÃO



FONTE: Choo (2003, p. 404).

Para Tarapanoff (2006), no contexto da GI a informação se refere a todos os tipos de recursos informacionais que se originam na produção de dados de diversas fontes organizacionais e são capazes de gerar valor para o negócio. Segundo a autora, a GI é a aplicação do ciclo de vida da informação nas organizações no sentido de identificar e potencializar recursos informacionais para o aprendizado e adaptação organizacional. A figura 3 ilustra o ciclo informacional mencionado pela autora.

FIGURA 3 – CICLO DE VIDA DA INFORMAÇÃO



FONTE: Tarapanoff (2006).

Sobre o ciclo de vida da informação, Tarapanoff discorre que

é iniciado quando se detecta uma necessidade informacional, um problema a ser resolvido, uma área ou assunto a ser analisado. É um processo que se inicia com a busca da solução a um problema, da necessidade de obter informações sobre algo, e passa pela identificação de quem gera o tipo de informação necessária, as fontes e o acesso, a seleção e aquisição, registro, representação, recuperação, análise e disseminação da informação, que, quando usada, aumenta o conhecimento individual e coletivo. (Tarapanoff, 2006, p. 23).

Nessa perspectiva, segundo a autora, a gestão da informação se manifesta ao gerar inteligência para a tomada de decisão, melhorando os processos e possibilitando eficiência organizacional, tendo, portanto, utilidade estratégica para a organização.

Por sua vez, Barbosa (2008) aborda a gestão da informação relacionando-a com a disciplina gestão do conhecimento (GC). Para o autor, GI é uma disciplina que se ocupa da documentação, da informação registrada, do conhecimento explícito; já a GC trata do conhecimento pessoal e tácito. Embora as duas disciplinas tenham suas particularidades, Barbosa (2008, p. 14) enfatiza que são fenômenos indissociáveis, ao passo que ambas “lidam com processos complexos, elusivos e de difícil observação” e “focalizam aspectos complementares de dois importantes fenômenos organizacionais”. Segundo o autor, a GI e a GC são processos difíceis de serem gerenciados e dependem totalmente da cultura informacional e organizacional, mas se configuram, cada vez mais, como fatores importantes nas transformações econômicas e sociais (Barbosa, 2008).

Com base nas definições apresentadas por Davenport, Choo, Tarapanoff, Barbosa e outros teóricos da GI, Nonato e Aganette propõem uma definição atual e consensual para o termo que sintetiza as conceituações acima apresentadas. Para Nonato e Aganette (2022), a gestão da informação é um

processo de gestão do ciclo de vida da informação que compreende as etapas: identificação das necessidades, criação, aquisição, organização, armazenamento, disseminação, distribuição e uso da informação. Pertence a uma organização, sofrendo ação direta da cultura organizacional e do planejamento estratégico. Através da análise dos fluxos da informação e da própria informação, visa à recuperação e a criação de produtos da informação. Utiliza as tecnologias da informação e comunicação, através dos sistemas de informação, como ferramenta de controle. (Nonato; Aganette, 2022, p. 148).

Conforme evidenciado pela literatura apresentada, percebe-se forte aplicação da GI nos ambientes organizacionais. Entretanto, é pertinente retratar também a

gestão da informação sob seu aspecto social. Para Araújo (1999), a interação humana pressupõe recepção, geração ou transferência de informação e, nesse sentido, toda prática social pode ser considerada uma prática informacional. Sob esse aspecto, a autora aponta que a relação entre informação e cidadania se caracteriza “pelo elemento diretamente político, ou seja, a informação pode auxiliar na construção de espaços sociais de cidadania” (Araújo, 1999, p. 166). A autora ainda aponta que a informação só será capaz de gerar transformações no tecido social se for gerada para a ação política de forma direta ou se for gerenciada de modo a identificar e atender necessidades específicas, seguindo princípios e processos da gestão da informação e do conhecimento.

Posto o referencial teórico apresentado nesta seção, a Gestão da Informação se manifesta na presente pesquisa ao fundamentar o campo de estudo para a manipulação, tratamento, mineração e análise dos dados, a fim de gerar informações relevantes que podem ser utilizadas para descoberta de conhecimento útil quanto ao fenômeno da violência de gênero no país. Desse modo, a próxima seção aborda outro conceito de suma importância para o entendimento do problema de pesquisa proposto na introdução, a análise de dados.

2.2 ANÁLISE DE DADOS

A análise de dados pode ser entendida como um processo que envolve a inspeção, limpeza, transformação e modelagem de dados para descobrir informações úteis, padrões, tendências e *insights* significativos. Esse processo é realizado com o intuito de embasar decisões através de informações relevantes, desenvolver estratégias eficazes, solucionar problemas e obter uma compreensão mais profunda sobre um determinado fenômeno ou contexto. O processo é desenvolvido por meio de algoritmos, ferramentas e métodos específicos, que podem incluir estatística descritiva, visualização de dados, mineração de dados, aprendizado de máquina e outras abordagens.

Para Sharda, Delen e Turban (2019), no livro onde abordam a análise de dados aplicada à gestão de negócios, ela pode ser vista como o processo de desenvolvimento de decisões ou recomendações práticas para ações baseadas em vislumbres gerados por dados históricos. Os autores ainda citam a análise de dados como representante da combinação entre tecnologia computadorizada, técnicas de

ciência administrativa e estatística direcionadas para a solução de problemas reais. De maneira complementar, apontam classificações para a análise de dados, a saber: descritiva, preditiva e prescritiva, explicadas abaixo.

Segundo os autores, a análise de dados descritiva trata de compreender e descrever o estado atual da organização, através da extração e análise de relatórios provenientes da consolidação de fontes e disponibilização de dados; a preditiva visa determinar probabilisticamente situações futuras, com base em técnicas estatísticas e mineração de dados; a prescritiva tem como meta “reconhecer o que está acontecendo, bem como o que deve vir a acontecer, e tomar decisões para garantir o melhor desempenho possível” (Sharda; Delen; Turban, 2019, p. 29).

Vale ressaltar que a análise de dados deve ser conduzida de forma rigorosa e imparcial, garantindo a validade e a confiabilidade das conclusões obtidas. Desse modo, a ética deve ser adotada em todas as etapas do processo de análise de dados, desde a definição dos objetivos da análise, coleta e preparação dos dados, exploração e transformação, até a aplicação de métodos analíticos e interpretação dos resultados.

Nesse sentido, ao abordar a análise de dados como uma sequência de etapas capaz de gerar informações valiosas, é possível entendê-la como parte intrínseca ao processo de descoberta de conhecimento em bases de dados, conhecido pela sigla KDD (*Knowledge Discovery in Databases*), assunto que será abordado na seção seguinte.

2.3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (KDD)

Steiner *et al.* (2006) definem a descoberta de conhecimento em bases de dados como uma área de estudo que objetiva abordar técnicas e ferramentas de transformação de dados armazenados através de atividades contínuas que visam a geração de conhecimento.

Fayyad, Piatetsky-Shapiro e Smyth (1996), ao relacionar as etapas e fatores que levam ao conhecimento, propuseram a descoberta de conhecimento em bases de dados como um processo que envolve a tríade dado-informação-conhecimento. Os autores descrevem o que chamaram de processo KDD como uma sequência de etapas que olha para o conhecimento desde sua gênese, na seleção da fonte de dados, passando pela geração de informações, através de técnicas de pré-processamento, transformação e mineração de dados, que, ao final, leva à descoberta

de conhecimento. O processo KDD é um processo interativo, porque permite ao usuário tomar decisões durante as etapas, e iterativo, porque é dado por refinamentos sucessivos, ou seja, as várias etapas do processo podem ser repetidas continuamente até que seja alcançado o conhecimento desejado (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

Goldschmidt, Passos e Bezerra (2015) apresentam duas definições para KDD: uma sob a perspectiva do conhecimento extraído e outra sob a perspectiva da realização do processo. Para a primeira perspectiva, Fayyad, Piatetsky-Shapiro e Smyth (1996 apud Goldschmidt; Passos; Bezerra, 2015, p. 3) definem KDD como “um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados”. Essa definição, segundo Goldschmidt, Passos e Bezerra (2015), enfoca as características do conhecimento resultante do processo de KDD, ressaltando que deve obedecer às premissas de ser válido, novo, útil, interessante e inteligível.

Por outro lado, os autores enfatizam que o KDD é um processo e limitar a atenção aos resultados pode levar à desconsideração da complexidade da extração, organização e apresentação do conhecimento extraído. Nesse sentido, sob a segunda perspectiva, Brachman e Anand² (1996 apud Goldschmidt; Passos; Bezerra, 2015) definem KDD como um processo que consiste em uma sequência de interações complexas, que acontece em um determinado período de tempo, entre um usuário e uma coleção de dados, frequentemente intermediado por um conjunto variado de ferramentas computacionais. Nessa perspectiva, o foco está nas etapas, técnicas e ferramentas do processo que leva ao conhecimento, envolvendo atividades como “negociação com os donos dos dados que se quer analisar, a estreita interação com os dados, e a manipulação de dados incompletos, inconsistentes ou incertos” (Goldschmidt; Passos; Bezerra, 2015, p. 6).

Em síntese, pode-se concluir que o processo KDD aplica conceitos de base de dados, métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial para utilização nas etapas de seleção, pré-processamento, transformação, mineração de dados e avaliação/interpretação dos resultados (Galvão; Marin, 2009).

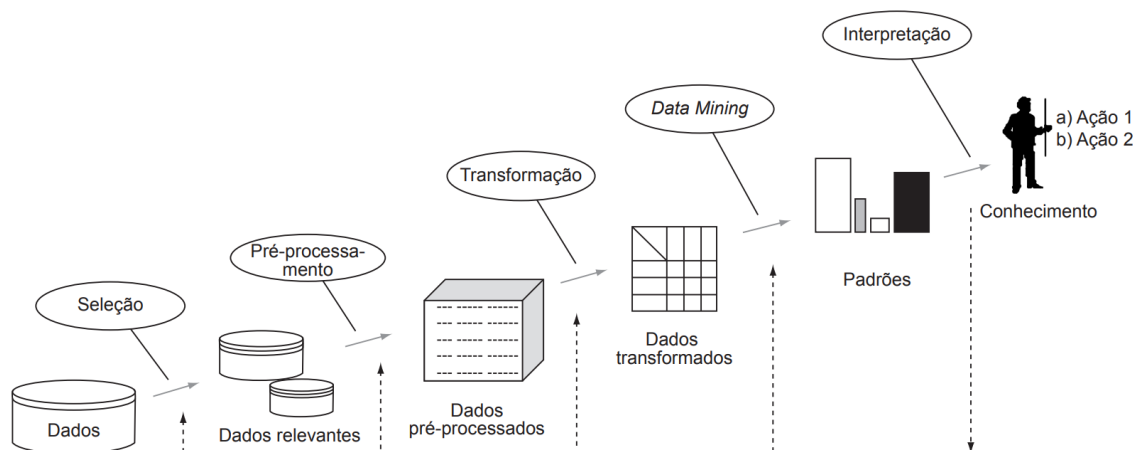
² BRACHMAN, R. J.; ANAND, T. The process of knowledge discovery in databases. The KDD process for extracting useful knowledge from volumes of data, 1996, p. 37-57

A proposta de Fayyad, Piatetsky-Shapiro e Smyth (1996) para o processo KDD tem como etapas principais a seleção dos dados, pré-processamento, transformação, mineração de dados e a interpretação das informações e conhecimento obtido, conforme evidencia a figura 4. Para descrever melhor o processo, os autores dividem essas etapas em nove passos, descritos a seguir:

1. compreensão do domínio de aplicação e do conhecimento prévio relevante para identificar o objetivo do processo KDD através do ponto de vista do cliente;
2. criação dos dados alvo: seleção de um conjunto de dados ou de subconjuntos de variáveis ou amostras de dados sobre o fenômeno que se deseja descobrir conhecimento;
3. limpeza e pré-processamento de dados, incluindo operações básicas como remoção de ruído e estratégias para lidar com dados ausentes;
4. redução da dimensionalidade, definindo os métodos mais efetivos para representação dos dados com número reduzido de variáveis, a depender do objetivo de descoberta;
5. definição dos métodos de mineração de dados que sejam apropriados para os objetivos do processo KDD, tais como sumarização, classificação, regressão, agrupamento e outros;
6. análise exploratória e seleção do algoritmo de mineração dos dados, definindo os modelos e parâmetros que sejam viáveis para a busca de padrões;
7. mineração de dados, buscando padrões de interesse através de regras de associação, árvores de decisão, regressão e agrupamento, considerando o usuário na escolha do método;
8. interpretação dos resultados, através de visualização dos padrões minerados e dos modelos extraídos, sendo possível iterar a partir de qualquer um dos passos anteriores quantas vezes sejam necessárias para melhorar a interpretação;
9. aplicação do conhecimento descoberto, utilizando-o para realização de ações, além de documentação deste para aplicações futuras, sendo possível então a verificação e a checagem de conhecimentos prévios.

Para os fins desta pesquisa, o processo KDD será abordado através de três etapas principais: pré-processamento, processamento e pós-processamento.

FIGURA 4 – PROCESSO KDD



FONTE: Steiner et al. (2006).

Para atingir o objetivo geral deste trabalho e realizar uma análise crítica sobre a violência contra a mulher, uma etapa imprescindível é a mineração de dados. Desse modo, a seção abaixo se ocupa em apresentar melhor esse assunto.

2.3.1 Mineração de dados

Pode-se dizer que a mineração de dados (comumente chamada *Data Mining* - DM - na literatura) é uma das principais etapas do processo KDD, pois é onde ocorre a busca efetiva por conhecimentos novos e úteis a partir dos dados (Goldschmidt; Passos; Bezerra, 2015, p. 65). Devido sua importância, a mineração de dados é frequentemente utilizada como sinônimo de KDD, porém cabe enfatizar que se trata apenas de uma etapa de um processo maior e consiste na aplicação de algoritmos de análise e descoberta que, sob limitações aceitáveis de eficiência computacional, produzem uma enumeração específica de padrões sobre os dados (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

Segundo Castro e Ferrari (2016), DM é uma etapa capaz de explorar uma base de dados pré-processados usando algoritmos adequados para a obtenção de conhecimento. Ainda conforme os autores,

é uma disciplina interdisciplinar e multidisciplinar que envolve conhecimento em áreas como banco de dados, estatística, aprendizagem de máquina, computação de alto desempenho, reconhecimento de padrões, computação natural, visualização de dados, recuperação de informação, processamento de imagens e de sinais, análise espacial de dados, inteligência artificial, entre outras. (Castro; Ferrari, 2016, p. 28).

Para Silva, Peres e Boscarioli (2016), a mineração de dados se trata de um processo que aplica técnicas e algoritmos computacionais para receber como entrada um conjunto de fatos ocorridos no mundo real e devolver como saída um padrão de comportamento, que pode ser expresso, por exemplo, como uma regra de associação, uma função de mapeamento ou a modelagem de um perfil.

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), DM envolve o ajuste de modelos ou a determinação de padrões a partir de dados observados. Ao se aprofundar sobre a importância desta etapa no KDD, os autores destacam que os objetivos da descoberta de conhecimento devem ser definidos pelo uso intencional do sistema. Para tanto, apontam que há dois objetivos principais no processo: a verificação e a descoberta. No primeiro objetivo, o sistema se limita em verificar a hipótese formulada pelo usuário. Já o segundo objetivo vai um pouco além e exige que o sistema encontre novos padrões de maneira autônoma.

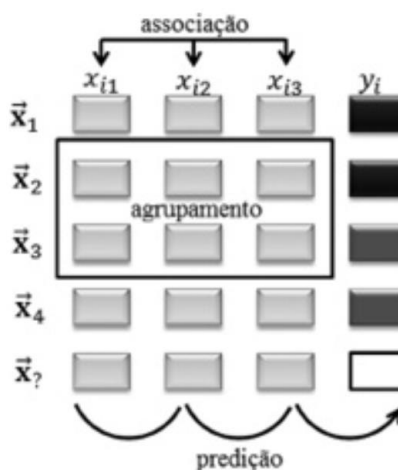
2.3.1.1 Tarefas de mineração de dados

Para Fayyad, Piatetsky-Shapiro e Smyth (1996), quando o objetivo do KDD é a descoberta, é possível dividir a mineração de dados em duas tarefas de alto nível, a saber: preditiva ou descritiva. As tarefas preditivas, segundo os autores, se ocupam em encontrar padrões para prever o comportamento futuro de algumas entidades. Quanto às tarefas descritivas, as definem como aquelas capazes de encontrar padrões para descrever os dados de maneira compreensível ao usuário. Os autores ainda salientam que, embora os limites entre predição e descrição não sejam nítidos, fazer uma distinção entre eles é útil para compreender o objetivo geral da descoberta.

Silva, Peres e Boscarioli (2016), buscando uma visão mais abrangente sobre as tarefas de mineração de dados, as enumeram em três grandes tarefas: predição, agrupamento e associação, conforme ilustrado na figura 5. Segundo os autores, as tarefas de predição têm como objetivo descobrir um modelo capaz de mapear corretamente cada um dos dados aos seus rótulos, o qual é alcançado através das chamadas técnicas supervisionadas, que são capazes de encontrar o modelo de

mapeamento a partir de procedimentos que associam um dado a um rótulo e corrigem tal associação quando ela não corresponde ao rótulo esperado (Silva; Peres; Boscarioli, 2016, p. 9). Os autores ainda dividem a análise preditiva em duas subtarefas: análise preditiva categórica, também chamada de tarefa de classificação; e análise preditiva numérica, também chamada de tarefa de regressão. A subtarefa de classificação se manifesta quando os rótulos associados aos dados pertencem a um conjunto discreto e finito de categorias. Por sua vez, a subtarefa de regressão se faz presente quando os rótulos associados aos dados são numéricos e pertencentes a um conjunto de valores contínuos.

FIGURA 5 – TAREFAS DE MINERAÇÃO DE DADOS



FONTE: Silva, Peres e Boscarioli (2016).

Quanto à tarefa de agrupamento de dados, Silva, Peres e Boscarioli (2016) explicam que consiste na análise de conjuntos de dados em que estão presentes apenas suas descrições, não havendo necessidade do uso de informação sobre qualquer tipo de rotulação. Essa tarefa objetiva descobrir relações entre os dados por meio de suas similaridades e fornecer, como resposta, a indicação de quais dados são similares entre si, oferecendo um modelo de agrupamento ou perfis para grupos de dados. Os algoritmos de agrupamento executam procedimentos que organizam os dados em grupos, de forma que a similaridade entre os dados de um grupo seja máxima e a similaridade entre dados colocados em grupos diferentes seja mínima (Silva; Peres; Boscarioli, 2016).

Por fim, Silva, Peres e Boscarioli (2016) definem a tarefa de associação como a busca por ocorrências frequentes e simultâneas entre elementos de um contexto,

onde os algoritmos que a resolvem analisam conjuntos de dados que representam eventos ou transações, procurando por itens frequentemente envolvidos nos mesmos eventos ou que apresentam algum tipo de correlação em seus comportamentos em tais eventos.

Essa pesquisa se valerá de métodos e algoritmos de mineração de dados a fim de gerar informações que podem ajudar na identificação de pontos críticos onde intervenções podem ser direcionadas, como programas de prevenção, conscientização e políticas públicas voltadas para a redução da violência contra a mulher, assunto abordado na próxima seção.

2.4 VIOLÊNCIA CONTRA A MULHER

A violência contra a mulher (VCM) é um fenômeno que atinge todas as classes sociais, etnias, religiões e culturas, ocorrendo em populações de diferentes níveis de desenvolvimento econômico e social (Ferraz, 2008). Tem como raiz os diferentes valores atribuídos culturalmente a mulheres e homens que determinam expectativas sobre seus comportamentos (FBSP, 2023).

Conforme Silva e Oliveira (2015), a violência sofrida pelas mulheres pode ser denominada também como violência doméstica (VD) ou violência de gênero e consiste em um fenômeno extremamente complexo, que atinge mulheres em todas as partes do mundo e tem suas raízes na inter-relação de fatores biológicos, econômicos, culturais, políticos e sociais. As autoras afirmam que o crescente número de estudos científicos sobre a VCM e, sobretudo, a veiculação das informações levantadas por essas pesquisas pode ajudar a compreender as circunstâncias nas quais o problema ocorre e contribuir para a prevenção e o enfrentamento do fenômeno, ao fornecer subsídios para políticas públicas que abranjam a saúde da mulher na sua integralidade.

Chauí³ (1985 apud Santos; Izumino, 2005) concebe a violência contra as mulheres como resultante de uma ideologia de dominação que define a condição feminina como inferior a masculina. Para a autora, as diferenças físicas e sociais entre

³ CHAUI, M. Participando do debate sobre mulher e violência. In: FRANCHETTO, B.; CAVALCANTI, M. L. V. C.; HEILBORN, M. L. (org.). **Perspectivas Antropológicas da Mulher 4**. Rio de Janeiro: Zahar, 1985.

feminino e masculino se transformam em desigualdades hierárquicas que, através de discursos ideológicos sobre o feminino, frequentemente incidem sobre o corpo da mulher.

Para Martinelli (2020), a violência contra a mulher é um fenômeno social global e se constitui como uma das formas de violação dos Direitos Humanos ao afetar meninas e mulheres em seu direito à vida e à saúde física e mental. A autora ressalta que, enquanto os homens são afetados em maior escala pela violência em ambientes públicos, é no ambiente privado que ocorre grande parte da violência contra as mulheres.

Nesse cenário, onde a VCM atinge a sociedade como um todo ao violar direitos humanos, causar impacto na saúde pública e no progresso econômico, o Brasil aprovou a Lei n.º 11.340/2006, uma importante política pública que cria mecanismos para coibir e prevenir a violência de gênero. Conhecida como Lei Maria da Penha, é uma das legislações mais avançadas mundialmente no tema e conceitua a violência doméstica e familiar contra a mulher como “qualquer ação ou omissão baseada no gênero que lhe cause morte, lesão, sofrimento físico, sexual ou psicológico e dano moral ou patrimonial” (Brasil, 2006).

Campos e Almeida (2017) defendem que este tipo específico de violência pode ser considerado como expressão máxima da desigualdade de gênero, fruto da estrutura social patriarcal que ordena e naturaliza as relações de poder entre homens e mulheres, produzindo e reproduzindo desigualdades sociais em diferentes espaços, perpetuando assim suas estruturas ao longo do tempo e em diversas sociedades, mesmo que de maneiras diferentes. Segundo os autores, o conceito de gênero foi inserido na produção acadêmica sobre as mulheres nos anos de 1970 e perpassa pela forma como a sociedade designa a homens e mulheres diferentes papéis, direitos e oportunidades de acordo com seu sexo (Campos; Almeida, 2017).

De maneira similar, Okabe e Fonseca (2008) argumentam que, para compreender o significado da violência de gênero, primeiramente é preciso conceituar gênero. Para tanto, abordam o gênero como uma construção sociológica e político-cultural do termo sexo, que se distingue da variável sexo por essa se referir à dimensão biológica da caracterização anátomo-fisiológica dos seres humanos. As autoras apontam o gênero como um “elemento constitutivo das relações sociais e como forma básica de representar as relações de poder, superando a visão de que as representações dominantes são naturais e inquestionáveis” (Okabe, Fonseca, 2008).

Segundo Santiago (2021), para que seja possível conhecer os fatores potencializadores e obstaculizadores da VCM e então planejar seu enfrentamento, o fenômeno precisa ser analisado em uma perspectiva ampla. A autora destaca a VCM como uma questão premente a ser priorizada pela gestão pública, enfatizando que não pode ser dissociada da desigualdade social e estrutural a qual está submetida boa parte das mulheres em situação de violência, caracterizada pela falta de acesso aos direitos sociais, desemprego, miséria, racismo, misoginia, entre outras expressões da problemática social.

Ao teorizar sobre a violência letal contra as mulheres, Lucena (2020), fundamentada pelas ideias de Walby^{4,5}, afirma que para entender as vulnerabilidades vividas pelas mulheres submetidas à violência torna-se fundamental perceber as dinâmicas atuais do patriarcado. A autora argumenta sobre a mutabilidade do regime de gênero, que, ao longo da história, se modificou de uma estrutura social patriarcal predominantemente privada, que confinava e explorava mulheres no trabalho doméstico, para uma estrutura social patriarcal pública, onde agora lhes é permitido o acesso, mas em condições de subordinação.

Lucena (2020) completa que as alterações nas dinâmicas das relações de gênero podem ser observadas em diferentes nuances: quanto à forma, do patriarcado privado para o público, e quanto ao grau de intensidade da opressão em dimensões específicas, como o mercado de trabalho e o Estado, que segregam mulheres a determinadas profissões e pagam menores salários pelo mesmo trabalho executado por homens. A autora acredita que a transição do patriarcado privado para o público modifica profundamente as situações de violência às quais as mulheres são expostas e, por consequência, as situações em que ocorrem os assassinatos quando elas são vítimas.

Ao observar a violência contra a mulher como a expressão máxima da desigualdade promovida pela sociedade patriarcal, o feminicídio se torna a expressão máxima da violência de gênero. Anjos Junior (2021) relembra que os homicídios de mulheres são menores comparativamente aos de homens, embora os casos dos crimes cometidos contra mulheres se destaquem, sobretudo, porque a principal motivação é devido ao gênero e em grande parte o agressor é alguém próximo à

⁴ WALBY, S. *Theorizing patriarchy*. Oxford: Basil Blackwell, 1990.

⁵ WALBY, S. *Gender transformations*. Londres: Routledge, 1997.

vítima. Nesse contexto, onde é possível perceber o feminicídio como a última etapa do ciclo de violência contra a mulher, a próxima seção se ocupa em definir o termo, visto que é um dos tipos de VCM analisado na presente pesquisa.

2.4.1 Feminicídio

De acordo com Miranda e Carvalho (2022, p. 8), “a linguagem é capaz de se adaptar aos novos anseios sociais e, assim, pode e deve ser modificada a partir de demandas que emergem do corpo social”. Nesse âmbito, o termo feminicídio se caracteriza como um neologismo e se refere ao assassinato de meninas e mulheres motivado pela misoginia, que é o ódio, preconceito ou desprezo pelo gênero feminino.

É derivado da conceituação do termo em inglês *femicide*, cunhado pela escritora e ativista feminista Diana Russell ao usá-lo pela primeira vez no Tribunal Internacional de Crimes contra Mulheres, em 1976, e caracterizá-lo como o assassinato de mulheres pelo fato de serem mulheres (Russell, 2011).

Mais tarde, Jill Radford lança, juntamente com Russell, o livro “*Femicide: the politics of woman killing*”, no qual as autoras associam o feminicídio a um *continuum* de violências sexuais praticadas contra mulheres e meninas que culminam em seu assassinato. Para as autoras, as raízes do feminicídio se entrelaçam com a origem do patriarcado, que perpetua nessa prática a sensação de poder e posse sob corpos femininos (Radford; Russell, 1992).

Embora a tradução literal de *femicide* seja femicídio, segundo explica Romio (2017), na América Latina a tradução do termo difundido por Russell se deu no contexto da luta pelos direitos das mulheres mexicanas na década de 1990, tendo como principal propagadora Marcela Lagarde, antropóloga, professora e feminista mexicana, que usou a palavra feminicídio para se referir a crimes que, além do autor, tem também responsabilização na negligência e ineficiência do Estado em promover políticas públicas para a proteção de meninas e mulheres (Lagarde, 2006).

O feminicídio expressa, de forma explícita, a desigualdade das relações entre homens e mulheres, nas quais a manifestação extrema de poder e terror são comportamentos que dão base para sua existência nas mais diversas sociedades (Miranda; Carvalho, 2022).

De acordo com a Lei brasileira nº 13.104/2015, feminicídio se configura como homicídio qualificado, constando no rol de crimes hediondos, e é caracterizado como

o assassinato de mulheres em decorrência de violência doméstica e familiar, menosprezo ou discriminação à condição de mulher (Brasil, 2015).

Para Pereira *et al.* (2022, p. 118), o feminicídio é um fenômeno social que se caracteriza como “uma modalidade de violência extremada consubstanciada por uma cultura de dominação e de inferiorização da condição das mulheres” presente em todas as sociedades. Os autores defendem que o feminicídio deve ser considerado um problema de saúde pública por sua natureza, extensão, severidade e significância à medida em que ceifa vidas, em sua maioria, na fase mais produtiva (Pereira *et al.*, 2022).

Conforme assume Romio (2017, p. 164), é impossível “identificar o nível máximo de feminicídios condensado em um indicador único, posto que não há uniformização na aplicação geral do conceito e consequente documentação dos casos”. Nesse contexto, a não observação da qualificadora de feminicídio em ocorrências acarreta em subnotificação e consequente perda de dados cruciais para o mapeamento do fenômeno.

Através da literatura consultada, identificou-se que é possível classificar o feminicídio sob diferentes perspectivas. Segundo Manjoo⁶ (2012 apud Caicedo-Roa; Bandeira; Cordeiro, 2022), ao assinar o relatório especial das Nações Unidas sobre violência contra as mulheres, o feminicídio se divide em duas categorias de análise: ativos, ou diretos, e passivos, ou indiretos. Sob esse aspecto, Romio (2017) pontua que no caso direto a autoria das mortes é masculina, enquanto o caso indireto tem como responsável o Estado, através de suas instituições e de grupos organizados. Ainda segundo a autora, com base na bibliografia analisada em sua tese,

o caso indireto tem a ver com mortes ligadas às políticas sexuais de controle do corpo feminino e reprodução (procriação), violência sexual sistemática no caso das guerras e conflitos, e omissão perante aos casos diretos. O caso direto é composto de mortes causadas pela violência, como o suicídio, homicídio e causas externas indefinidas, não correspondem ao total destas mortes, mas representam grande parte delas. (ROMIO, 2017, p. 70)

Para Caicedo-Roa, Bandeira e Cordeiro (2022), distinguir o feminicídio de outros tipos de homicídios femininos e elaborar uma tipologia para caracterizar os

⁶ MANJOO, Rashida. Report of the Special Rapporteur on violence against women, its causes and consequences. United Nations A/HRC/20/16.

diferentes tipos de feminicídio viabiliza a obtenção de dados específicos e precisos sobre esse fenômeno. Os autores apresentam a classificação de feminicídios de acordo com o modelo do protocolo latino-americano de Sarmiento⁷ *et al.* (2014 apud Caicedo-Roa; Bandeira; Cordeiro, 2022), que possui 13 categorias diferentes, a saber: feminicídio íntimo, não íntimo, infantil, familiar, por conexão, sexual sistêmico, por prostituição ou ocupações estigmatizadas, por tráfico de pessoas, por contrabando de pessoas, transfóbico, lesbofóbico, racista e por mutilação genital feminina.

Em sua tese, após analisar as diferentes categorizações existentes na literatura, Romio (2017) propõe investigar o feminicídio através da seguinte tipologia:

- Feminicídio reprodutivo: vinculado às políticas de controle do corpo e da sexualidade da mulher, cuja expressão mínima são as mortes por aborto registradas nos sistemas de saúde;
- Feminicídio doméstico: violência letal por agressão física contra a mulher no contexto domiciliar, conjugal ou familiar registrados no sistema de saúde;
- Feminicídio sexual: violência letal por agressão física contra a mulher por meio sexual.

Nesse contexto, em que o feminicídio se configura através da violência de gênero decorrente de relações desiguais de poder, se faz necessário destacar a relevância da utilização do termo para se referir ao assassinato de meninas e mulheres motivada pela misoginia. Conforme aponta Roichman (2020),

é de suma importância a tipificação do feminicídio para reconhecer, no arcabouço legislativo estatal, que mulheres estão morrendo pela simples razão de serem mulheres, expondo a fratura da desigualdade de gênero que ainda persiste na sociedade. Por outro lado, a tipificação evita que assassinos misóginos possam ser beneficiados com interpretações jurídicas anacrônicas e moralmente inaceitáveis, tendentes a amenizar o comportamento criminoso e sua punição, como a de prática de “crime passionai”. (Roichman, 2020, p. 359).

Ao analisar os efeitos da Lei nº 13.104/2015 e constatar aumento no número de assassinatos de mulheres desde sua aprovação, o autor enfatiza que a importância de tipificar especificamente o crime de feminicídio sobrepõe o impacto causado nos

⁷ SARMIENTO, C. et al. “Latin American model protocol for the investigation of gender-related killings of women (femicide/feminicide)”. Un Women. United Nations, 2014.

índices de mortalidade de mulheres, pois significa, em verdade, o reconhecimento da existência da prática e a oposição estatal à conduta criminalizada (Roichman, 2020).

2.5 RELAÇÕES ENTRE ANÁLISE DE DADOS E VIOLÊNCIA CONTRA A MULHER

A fim de demonstrar a pertinência de pesquisas que visam analisar dados sobre a violência de gênero, nesta seção serão apresentados estudos que relacionam ambos os assuntos.

Na dissertação intitulada “Proposta de fluxo de análise de dados preditiva em violência doméstica contra a mulher”, apresentada ao Programa de Pós-Graduação em Gestão da Informação da UFPR, Costa (2022) aplica métodos de KDD nas bases de dados Boletim de Ocorrência Unificado e Sistema de Ocorrências Letais, fornecidas pela Secretaria de Segurança Pública do Paraná, que contém registros de mulheres vítimas de violência doméstica e feminicídio dos anos 2018, 2019 e 2020.

No estudo, foram utilizadas ferramentas para tratamento, mineração e análise de dados como o Microsoft Excel e os *softwares* RStudio, WEKA e Orange. Na fase de pré-processamento dos dados, Costa (2022) realizou discretização e balanceamento por reamostragem, para então aplicar as técnicas CFS, Gain Ratio, Info Gain e ReliefF a fim de selecionar os atributos mais relevantes para a mineração dos dados a serem utilizados em análises preditivas para identificação de vítimas fatais de violência doméstica. Quanto à etapa de mineração, Costa (2022) optou por analisar as bases sob a ótica da estatística descritiva e dos algoritmos de indução de regras de classificação PRISM e CN2. Para a avaliação do modelo preditivo, o autor se valeu de métricas como acurácia e F1-Score, além de matrizes de confusão.

Através da execução das etapas de KDD, utilizando algoritmos e ferramentas de ciência de dados, além dos conhecimentos providos pelo desenvolvimento do estudo, Costa (2022) pode constatar que, para abarcar os diferentes tipos e propósitos de pesquisas futuras a serem apoiadas por seu trabalho, o produto da pesquisa deveria ser a proposição de não apenas um, mas sim três fluxos de análise preditiva para dados de violência doméstica, a saber: fluxo de análise de dados preditiva com estatística descritiva, fluxo de análise de dados preditiva com KDD e fluxo de análise de dados preditiva com KDD para base com reincidências.

O segundo estudo selecionado é a tese “Feminicídios no Brasil, uma proposta de análise com dados do setor de saúde”, onde Romio (2017) aborda o feminicídio

sob o olhar da demografia, através da articulação entre teoria feminista e análise empírica da mortalidade feminina, com base em dados de registros do Sistema Único de Saúde (SUS) entre os anos 1996 a 2014.

De natureza empírica, o estudo de Romio (2017) buscou caracterizar demograficamente o feminicídio ao analisar três fontes de informação do SUS, a saber: autorizações de internação hospitalares, declarações de óbito e fichas de notificação compulsória da violência sexual e doméstica contra mulheres. Nas duas primeiras bases mencionadas, a autora comparou estatisticamente dados dos gêneros masculino e feminino, a fim de identificar diferencial de gênero na mortalidade sob aspectos como faixa etária e agressão sexual. A terceira base foi analisada a partir da segmentação dos dados em três categorias: feminicídio reprodutivo, onde a autora selecionou os dados referentes a mortes por aborto; doméstico, contendo dados sobre mortes por agressão ocorridas na residência da vítima e/ou no contexto familiar e conjugal; e o feminicídio sexual, onde foram consideradas as mortes decorrentes de agressão por meio sexual.

Romio (2017) esclarece que a escolha por dados do setor de saúde, em detrimento de dados de bases criminais, se deu pelo fato de essa fonte de informação ser de acesso público e contar com sistematização e cobertura nacional, em contraponto aos registros criminais, que não possuem um sistema de informação nacional unificado e têm o acesso regulado pela Segurança Pública dos estados. Nesse sentido, Romio (2017) explica que, para analisar dados sobre feminicídio provenientes de sistemas de saúde, considera-se como critério de seleção dos dados a Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10), visto que, embora seja tipificado juridicamente no Brasil, não há ainda categorização em documentos médicos para o feminicídio.

Na literatura sobre análise em dados dos sistemas de informação de saúde é consenso associar o feminicídio ao capítulo XX do CID-10; este capítulo categoriza causas externas de morbidade e mortalidade e compreende as categorias V01 a Y98. Autores como Waiselfisz (2015), Meneghel e Hirakata (2011) e Scherer *et al.* (2022), ao analisar dados provenientes do Sistema de Informações sobre Mortalidade (SIM), optaram pelos dados correspondentes às categorias X85 a Y09 do CID-10, que versam sobre mortes por agressão, para caracterizar o feminicídio. Romio (2017), por outro lado, a fim de realizar uma análise mais abrangente do fenômeno, analisou também dados de diferentes capítulos da CID-10, como o XV, que versa sobre

gravidez, parto e puerpério, para incluir na análise os feminicídios reprodutivos, os quais a autora aponta como feminicídios indiretos.

Romio (2017) conclui a tese enfatizando que, embora cada uma das bases selecionadas produza indicadores próprios sobre o fenômeno, visto as diferenças nas suas naturezas, analisá-las no mesmo estudo evidenciou as limitações de fontes de dados do setor de saúde, pois estas não foram elaboradas para capturar dados sobre o feminicídio e carecem de dados sobre os autores e causas presumíveis do fato.

Ainda no contexto de fontes de dados dos sistemas de saúde, Meneghel e Hirakata (2011), em sua pesquisa intitulada “Femicídios: homicídios femininos no Brasil”, inicialmente pretendiam analisar especificamente femicídios (termo que também era usado à época), mas perceberam que essa informação não era identificada nas declarações de óbito, e, portanto, não constava no Sistema de Informações sobre Mortalidade. Desse modo, as autoras optaram por “utilizar o total de homicídios femininos como um ‘indicador aproximado’ de feminicídio, já que de 60% a 70% dessas mortes são provocadas pelas desigualdades de gênero” (Meneghel; Hirakata, 2011, p. 566). Para tanto, analisaram dados provenientes do SIM sobre as mortes femininas ocorridas por agressão entre 2003 e 2007, classificadas nas categorias X85 a Y09 do CID-10.

Vale lembrar que, à época do artigo, os homicídios femininos pela condição de gênero ainda não eram tipificados juridicamente, fato que só foi possível após a aprovação da Lei 13.104, em 2015. Nesse sentido, Meneghel e Hirakata (2011) apontam que usar o total das mortes femininas por agressão como *proxy* dos feminicídios na população brasileira poderia causar a superestimação dos valores reais do evento; por outro lado, ponderam que tal abordagem possa ser compensada pela subnotificação ou diagnóstico mal definido de agressão a mulheres, fatos que eram ainda mais comuns àquela época do que são hoje.

As autoras caracterizam o artigo como de natureza ecológica, pois buscam relacionar a mortalidade feminina por agressão com variáveis socioeconômicas, demográficas e de saúde. Para tanto, utilizam regressão linear a fim de correlacionar as variáveis já mencionadas com o coeficiente de mortalidade feminina por agressão, definido como variável dependente, que foi padronizado através do valor populacional padrão entre 2000 e 2025 estimado pela Organização Mundial da Saúde. Meneghel e Hirakata (2011) explicam que os coeficientes de mortalidade foram padronizados para

que fosse possível efetuar comparações em regiões com diferentes padrões demográficos e etários, como é o caso das regiões brasileiras.

3 ENCAMINHAMENTOS METODOLÓGICOS

Esta seção se dedica a apresentar elementos inerentes ao desenvolvimento deste trabalho, como a caracterização da pesquisa, os materiais utilizados e métodos aplicados.

3.1 CARACTERIZAÇÃO DA PESQUISA

A caracterização deste estudo se baseia na classificação de Sampieri, Collado e Lucio (2013) quanto ao enfoque da abordagem, ao alcance e à estratégia, a qual os autores denominam “desenho de pesquisa”.

Segundo Sampieri, Collado e Lucio (2013), a abordagem de pesquisa pode ter enfoque qualitativo ou quantitativo. O presente estudo pode ser classificado como quantitativo, pois essa abordagem, segundo os autores, “utiliza a coleta de dados para testar hipóteses, baseando-se na medição numérica e na análise estatística para estabelecer padrões e comprovar teorias” (Sampieri; Collado; Lucio, 2013, p. 30).

Quanto ao alcance da pesquisa, o que os autores definem como um contínuo de causalidade que um estudo pode ter, Sampieri, Collado e Lucio (2013) categorizam em pesquisa exploratória, descritiva, correlacional ou explicativa. A presente pesquisa pode ser considerada como exploratória, pois estudos que se enquadram nessa categoria são desenvolvidos com o objetivo de examinar um tema de pesquisa pouco estudado, sobre o qual ainda se tem muitas dúvidas ou se deseja pesquisar sob novas perspectivas, ou um tema que ainda não foi abordado cientificamente (Sampieri; Collado; Lucio, 2013, p. 100).

Por fim, quanto à estratégia, Sampieri, Collado e Lucio (2013) classificam a pesquisa quantitativa em experimental ou não-experimental. Nesse sentido, a presente pesquisa pode ser classificada como pesquisa experimental, porque se trata de uma situação de controle na qual se manipula, de maneira intencional, uma ou mais variáveis independentes (causas) para analisar as consequências dessa manipulação sobre uma ou mais variáveis dependentes (efeitos) (Sampieri; Collado; Lucio, 2013, p. 142).

3.2 MATERIAIS E MÉTODOS

Nesta pesquisa, o processo KDD apresentado na segunda seção é aplicado em três etapas principais - pré-processamento, processamento e pós-processamento - as quais se relacionam diretamente com os objetivos específicos informados na primeira seção. De modo a viabilizar a execução de cada uma dessas etapas, se faz necessária a utilização de ferramentas tecnológicas como *softwares* e linguagem de programação. A seguir, são apresentadas as ferramentas utilizadas para o desenvolvimento dessa pesquisa.

Para compreensão inicial quanto à estrutura e dimensionalidade das bases fornecidas pelas Secretarias de Segurança Pública dos estados Pará e Paraná, bem como o conteúdo das variáveis e tipos dos dados, utilizou-se a ferramenta Microsoft Excel, um *software* de edição de planilhas que possibilita a organização e visualização dos dados através da construção de tabelas, além de ser uma poderosa ferramenta de cálculo. Rapidamente pode-se perceber que os tratamentos que deveriam ser realizados nos conjuntos de dados não seriam suportados de maneira satisfatória pelo Excel, visto o tamanho dos *datasets*, com alguns passando dos 300.000 registros.

Nesse sentido, planejou-se a estratégia de realizar o tratamento e análise descritiva dos dados através de uma linguagem de programação. A selecionada foi a linguagem Python, que é uma linguagem de programação interpretada de alto nível, fortemente e dinamicamente tipada (Lambert, 2022; Netto; Maciel, 2021). A escolha do Python se deu por conhecimento prévio da autora sobre a linguagem, aprendida através da disciplina Programação de Computadores para Gestão da Informação, além do fato de ser uma linguagem amplamente utilizada no mundo dos dados, com suporte a diversas demandas de tratamento, análise e *machine learning*.

Para aplicação do Python em conjuntos de dados é comum utilizar *frameworks* e bibliotecas, que são coleções de funções e subprogramas desenvolvidos para demandas específicas. Nesta pesquisa, utilizou-se PySpark, que é uma API que combina a facilidade de aprendizado do Python com o poder de processamento para tratamento, análise de dados e aprendizado de máquina do *framework* Apache Spark. Do PySpark, foi utilizado o módulo `pyspark.sql`, que permite a manipulação de dados estruturados através de sintaxe muito similar à linguagem de consulta SQL, e `pyspark.ml`, um pacote de aprendizado de máquina dimensionável capaz de fornecer APIs de alto nível para a criação de *pipelines* e modelos de *machine*

learning. Do módulo `pyspark.sql`, foram utilizadas as classes `functions`, `types` e `window`. Da classe `functions`, usou-se as funções `col`, `lower`, `when`, `substring`, `hour`, `to_timestamp`, `date_format`, `count`, `current_date`, `datediff`, `mean`, `median`, `variance`, `stddev`, `asc`, `expr`, `lit`, `split`, `upper`, `avg` e `regexp_replace`, que permitem a manipulação de colunas, string e tipos de dados, além de agregação de valores e cálculo de medidas estatísticas.

Conforme descrito na figura 6, também foram utilizadas as bibliotecas Google.Colab, Numpy, Shutil, Pandas, Matplotlib, Seaborn, Folium e Branca sendo essas quatro últimas para a criação dos gráficos presentes nessa pesquisa.

FIGURA 6 – INSTALAÇÃO DE BIBLIOTECAS, MÓDULOS E PACOTES

```
#importando módulos e pacotes
from pyspark.sql import SparkSession
from pyspark.sql import functions as f
from pyspark.sql.types import StringType, StructType, StructField, DateType
from pyspark.ml import stat
from pyspark.sql.window import Window
from folium.plugins import HeatMap
from google.colab import drive
import matplotlib.pyplot as plt
import pandas as pd
import shutil
import folium
import branca.colormap
import seaborn as sns
```

FONTE: A autora (2023).

Para o desenvolvimento dos códigos de tratamento e análise descritiva dos dados, utilizou-se o Google Colab, um serviço de Jupyter Notebook hospedado em nuvem, que visa o incentivo à pesquisa em *machine learning*. O Colab foi o editor e interpretador de código escolhido por suportar o PySpark, que tem como pré-requisito sistemas operacionais Linux e, portanto, não pode ser executado localmente no sistema operacional Windows utilizado para o desenvolvimento deste trabalho, além da facilidade de acesso provido pela nuvem Google.

Para a mineração dos dados, modelagem de aprendizado de máquina e avaliação dos modelos foi utilizado o *software* Orange Data Mining, uma ferramenta *open source* que possui interface gráfica e utilização *user-friendly*, que suporta desde a importação de dados até a modelagem de *machine learning*, com ênfase na avaliação de modelos. Coneglian *et al* (2023) destacam o Orange como uma

ferramenta valiosa para a inserção da Inteligência Artificial no desenvolvimento de pesquisas das Ciências Sociais Aplicadas, permitindo uma compreensão mais profunda e precisa de fenômenos sociais e comportamento humano através de sua interface intuitiva, flexibilidade de personalização e recursos avançados.

3.2.1 Base de dados

Esta seção se ocupa em apresentar a descrição da base de dados construída para a análise desta pesquisa. Trata-se da união de duas bases distintas, provenientes dos registros criminais de violência doméstica e feminicídio ocorridos entre os anos 2018 e 2022, fornecidos pelas Secretarias de Segurança Pública do Pará e do Paraná.

A base paraense foi disponibilizada em formato .xlsx e tem 196.256 registros e 68 atributos. São ocorrências registradas pelas delegacias do estado sobre os crimes de violência doméstica em geral, e, após a aplicação de todos os filtros descritos detalhadamente na próxima seção, contém 6.896 linhas que se referem especificamente à violência contra a mulher.

A base do Paraná foi disponibilizada em quatro arquivos diferentes, todos em formato .csv, que são complementares. Um arquivo se refere ao registro da ocorrência, com dados como o código identificador do boletim de ocorrência, data e hora do fato e do registro do boletim, bem como o município e o bairro e conta com 340.478 linhas e 18 atributos. Outro arquivo contém as naturezas dos crimes, sendo que os mesmos atributos relatados na base de ocorrências também estão presentes, mas nesse outro arquivo há dados adicionais como o local de ocorrência do crime, o meio empregado e o tipo de crime cometido. A base de naturezas tem 468.441 linhas e 22 atributos, sendo que há 340.478 valores únicos para o identificador do boletim de ocorrência, o que indica que o mesmo boletim pode ter mais de um meio empregado ou especificação criminal. Segundo Costa (2022), é obrigatório para cada boletim o cadastro de pelo menos um meio empregado, porém é comum que em ocorrências de violência doméstica existam pelo menos dois valores cadastrados.

O terceiro arquivo referente à base do Paraná é o que contém dados sobre as vítimas. Em conformidade com a Lei Geral de Proteção de Dados, essa base já foi disponibilizada pela Secretaria de Segurança Pública do estado do Paraná sem dados sensíveis que identifiquem a vítima. São descritas suas características demográficas

como idade, faixa etária, cor da pele, grau de instrução e profissão. Há 350.240 linhas e 26 atributos, e, como na base naturezas, são 340.478 valores únicos de identificador do boletim de ocorrência, indicando que o mesmo boletim pode conter mais de uma vítima envolvida. O último arquivo é a base que contém as características demográficas do autor do crime, como idade, faixa etária, cor da pele, grau de instrução, profissão e grau de relacionamento com a vítima. Neste arquivo, há 10.400 linhas e 27 atributos, o que indica que para a grande maioria de ocorrências, não há informações disponibilizadas sobre o autor registradas no boletim de ocorrência.

Após todos os tratamentos realizados individualmente nas bases de cada estado e união delas em apenas um *dataset*, a base de ocorrências de crimes contra a mulher desta pesquisa conta com 24.534 registros, distribuídos entre os anos de 2018 a 2022, e 33 atributos, descritos no quadro 3, que foram definidos por similaridade entre as duas bases. Em ambas as bases foram disponibilizados apenas os registros, não sendo acompanhados por um dicionário de dados oficial; portanto, a descrição dos atributos foi inferida pela autora através de análise exploratória dos dados na ferramenta Microsoft Excel e conhecimentos adquiridos através da pesquisa.

QUADRO 3 – ATRIBUTOS DA BASE DE VIOLÊNCIA CONTRA A MULHER

Atributo	Tipo	Descrição	Exemplo
data_registro	Ordinal	Data de registro da ocorrência	16/10/2018
hora_registro	Ordinal	Hora de registro da ocorrência	08:38
data_fato	Ordinal	Data do fato	16/10/2018
hora_fato	Ordinal	Hora do fato	16:30
faixa_horaria_2h	Ordinal	Faixa horário do fato (a cada 2h)	22h - 00h
faixa_horaria	Ordinal	Período do dia em que ocorreu o fato	noite
dia_semana	Ordinal	Dia da semana em que ocorreu o fato	dom
mes_registro	Ordinal	Mês de registro da ocorrência	10
mes_fato	Ordinal	Mês do fato	10
ano_registro	Ordinal	Ano de registro da ocorrência	2018
ano_fato	Ordinal	Ano do fato	2018
município	Nominal	Município de ocorrência do fato	belem
bairro	Nominal	Bairro de ocorrência do fato	centro
estado	Nominal	Estado onde ocorreu o crime	pa
latitude	Contínuo	Latitude onde ocorreu o crime	-1,3747476
longitude	Contínuo	Longitude onde ocorreu o crime	-48,4165881

local_ocorrencia	Nominal	Local onde aconteceu o crime	residencia
meio_empregado	Nominal	Meio empregado no crime	armas em geral
classe_motivo	Nominal	Classificação do motivo segundo a legislação	crimes de violencia domestica - lei 11.340/2006
registros	Nominal	Modalidade de crime	lesao corporal
vit_dt_nasc	Ordinal	Data de nascimento da vítima	30/01/1979
vit_idade	Discreto	Idade da vítima	29
vit_faixa_etaria	Ordinal	Faixa etária da vítima	adulto ii (25 a 29 anos)
vit_cor_pele	Nominal	Cor da pele da vítima	preta
vit_grau_inst	Nominal	Grau de instrução da vítima	ensino fundamental incompleto
vit_profissao	Nominal	Profissão da vítima	dona de casa
aut_data_nasc	Ordinal	Data de nascimento do autor	06/07/1995
aut_idade	Discreto	Idade do autor	33
aut_faixa_etaria	Ordinal	Faixa etária do autor	adulto iii (30 a 34 anos)
aut_cor_pele	Nominal	Cor da pele do autor	branca
aut_grau_inst	Nominal	Grau de instrução do autor	ensino fundamental incompleto
aut_profissao	Nominal	Profissão do autor	motorista
grau_de_relaciona mento	Nominal	Grau de relacionamento e vítima e autor	ex-namorado(a)

FONTE: A autora (2023).

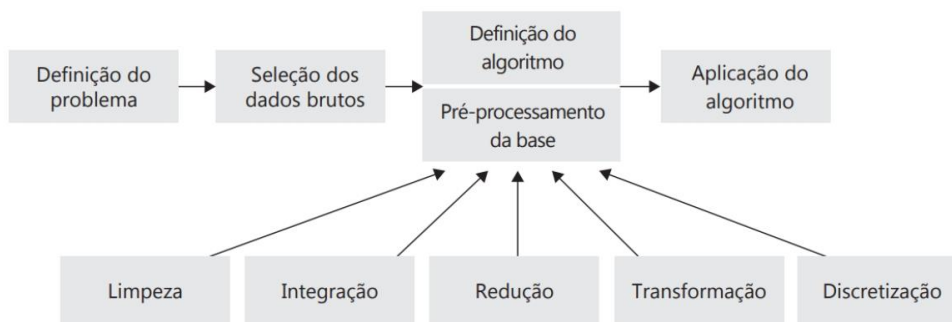
As próximas seções se ocupam em descrever os métodos utilizados para aplicação do processo KDD nesta pesquisa, que, embora tenha apenas três etapas principais, contempla todos os passos necessários para percorrer o caminho da descoberta de conhecimento da base de dados de violência contra a mulher.

3.2.2 Pré-processamento

Nesta seção, são contemplados os passos do processo KDD de compreensão inicial quanto aos objetivos da aplicação, criação dos dados alvo, limpeza e pré-processamento dos dados e de redução da dimensionalidade. Conforme Castro e Ferrari (2016), é comum que a etapa de pré-processamento consuma mais tempo de análise do que as etapas de aplicação dos algoritmos de mineração propriamente ditos, fato que se fez presente nesta pesquisa. Os autores afirmam que cada técnica de pré-processamento deve ser aplicada considerando seu efeito resultante na base de dados. Desse modo, defendem que entender a função e a aplicabilidade de cada método de pré-processamento é tão importante quanto entender como cada método

funciona. A figura 7 ilustra o processo de preparação da base de dados apresentado pelos autores.

FIGURA 7 – PROCESSO DE PREPARAÇÃO DA BASE DE DADOS



FONTE: Castro e Ferrari (2016).

Em seguida, são descritos os procedimentos aplicados em cada base individualmente, o processo de união delas para que fosse possível construir a base unificada e os tratamentos realizados também na base única. A manipulação de dados aplicada nesta e nas próximas duas seções foi desenvolvida em PySpark, utilizando o Google Colab. Para iniciar o ambiente do Colab habilitado para desenvolvimento em PySpark é necessário instalar o pacote através do comando pip, utilizado para gerenciamento de pacotes Python, importar a biblioteca que gera a conexão e criar uma sessão do Spark no *notebook*, conforme descrito na figura 8.

FIGURA 8 – INSTALAÇÃO DO PYSPARK E CONFIGURAÇÃO DO AMBIENTE

```

#instalando spark
!pip install pyspark

from pyspark.sql import SparkSession

#criando sessao spark
spark = SparkSession.builder\
    .master('local[*]')\
    .appName("transformacao_VCM")\
    .getOrCreate()
  
```

FONTE: A autora (2023).

Após a instalação do pacote PySpark, das bibliotecas e dos módulos supracitados na figura 6, a primeira base de dados tratada foi a do estado do Pará. A base possui originalmente 196.256 registros e 68 atributos. Após importar o arquivo .xlsx no Colab através da biblioteca 'drive' do pacote google.colab, e atribuir ele ao *dataset* 'dados_VCM_PA', aplicou-se no *dataframe* resultante um filtro na variável 'vit_sexo', pois, conforme citado anteriormente, a base paraense original contém as ocorrências de violência doméstica em geral do estado, incluindo a violência infantil, contra idosos, entre outras. Portanto, com o método .filter (), foram filtrados apenas os registros referentes a vítimas do gênero feminino, representadas pelo valor 'F' do atributo. Essa filtragem excluiu do conjunto de dados 25.190 registros, que continham valores como 'M', para o gênero masculino, 'P', de prejudicado (uma categoria presente em muitos atributos da base), ou os valores nulos.

Após o filtro para o gênero da vítima, foi aplicado um filtro no atributo 'aut_sexo', para excluir do *dataset* registros contendo autoras do gênero feminino, visto que para se enquadrar como VCM o crime deve ser cometido por autor do gênero masculino. Esse filtro excluiu do *dataset* 11.808 registros. O próximo filtro aplicado foi no atributo 'ano_fato', de modo a considerar apenas as ocorrências do ano 2018 em diante. Esse filtro se fez necessário pois a base paraense original contém ocorrências desde o ano 2010 até 2022 e a base do Paraná desde o ano 2018 até 2023; portanto, para união e comparação entre as bases, foi definido o período comum de 2018 a 2022. O filtro de ano removeu do *dataset* 64.837 registros.

O último filtro na base do Pará foi com o intuito de reduzi-la a ocorrências que sejam referentes somente à VCM, aplicando ao atributo 'especializacao_fato' para considerar apenas os registros onde os valores contêm as palavras-chave "mulher", "penha", "feminicídio", "femicídio", "feminicidio" ou "femicidio", descritores esses que foram definidos após análise prévia dos valores possíveis para esse atributo. Esse filtro removeu 87.525 linhas do *dataset*, que ao final ficou com 6.896 registros.

Posteriormente, alguns atributos foram desconsiderados por não apresentarem relevância para a análise. A decisão pela remoção desses atributos se deu pelo fato de conterem mais registros nulos ou inválidos do que dados válidos, além de haver necessidade de conciliar os atributos entre as bases de cada estado, pois a base paranaense original contém menos atributos que a paraense. A remoção dessas colunas, utilizando o método .drop(), resultou em um *dataset* com 32 colunas para a base do estado do Pará.

Após verificar o tipo dos dados através do método `.printSchema()`, percebeu-se que havia necessidade de tipar corretamente os atributos de data, como a data do fato, do registro, de nascimento da vítima e do autor, que foram tipadas automaticamente como *string*, para o tipo *date*. O próximo tratamento foi nos atributos 'hora_fato' e 'hora_registro', que tinham o tipo *timestamp*, com valores como '1970-01-01 16:47:00'; então, com o intuito de ficar apenas com a hora desse campo, foram extraídos somente os últimos 8 caracteres, procedimento que resulta em valores do tipo *string*.

Os atributos relativos a idades e faixas etárias, tanto da vítima, quanto do autor, continham erros de digitação em um número considerável de registros. Portanto, inicialmente, foram substituídos os valores já existentes dos atributos de idade pelo resultado da subtração entre a data do fato e a data de nascimento. Depois disso, com os novos valores para a idade, foi feita a reclassificação em faixas etárias. Ambas substituições foram feitas utilizando o método `.when()`. Após verificação das idades resultantes, foi possível identificar que haviam valores muito altos, como 120 anos. Tais idades foram calculadas pois havia considerável quantidade de registros de data de nascimento com o valor '1900-09-09'. Desse modo, as idades em que o ano de nascimento continha esse valor, ou data anterior a essa, foram substituídas pela categoria 'Desconhecida'.

Também foram feitas substituições nos atributos 'mes_fato' e 'mes_registro', onde os valores, que eram referentes ao mês por extenso, como 'NOVEMBRO', foram substituídos pelos números correspondentes, visto que na base do Paraná os valores dos meses são representados pelos números, além do fato de facilitar a ordenação correta para o agrupamento da contagem de cada mês utilizado na estatística descritiva.

Os próximos tratamentos no *dataset* do estado do Pará foram feitos com base no atributo 'hora_fato', com o intuito de substituir os valores registrados das faixas horárias do acontecimento do crime, que continham valores como '18 |-- 24', para a faixa horária a cada 6 horas, e '22 |-- 24', para a faixa horária a cada 2 horas. Utilizando o método `.when()`, a hora de acontecimento do crime foi classificada em faixas horárias com valores como 'Manhã', 'Tarde', 'Noite' e 'Madrugada', para a faixa horária a cada 6 horas, e '20h – 22h', para a faixa horária a cada 2 horas.

Os últimos tratamentos feitos na base do Pará foram a reordenação de colunas, de modo a deixar o *dataset* organizado de acordo com uma sequência lógica

de atributos, além de criar uma nova coluna com o método `.lit()` para identificar o estado a qual os dados pertencem com valor 'PA', para que, após o tratamento da base paranaense, possa ser feita a união de ambos os *datasets*. O *dataset* final dos crimes de violência contra a mulher do estado do Pará tem 6.896 registros e 33 atributos.

Devido a base do estado do Paraná ter sido disponibilizada em quatro arquivos diferentes, conforme citado anteriormente, foi adotada uma estratégia de tratamento de dados distinta da aplicada na base do Pará. Após exploração inicial dos dados, definiu-se que a base de naturezas seria utilizada como tabela principal, pois através dela é que seria possível distinguir os casos de VCM dos outros casos de violência doméstica.

Tal distinção foi feita através da filtragem, pelo método `.filter()`, do atributo 'classe_motivo', para que fossem selecionados apenas os registros contendo as palavras-chave "mulher", "sexual", "medidas protetivas", "13772". A última palavra-chave se refere à lei 13.772/2018, que altera a Lei Maria da Penha para reconhecer que a violação da intimidade da mulher configura violência doméstica e familiar e criminalizar o registro não autorizado de conteúdo com cena de nudez ou ato sexual de caráter íntimo e privado (Brasil, 2018). Esse procedimento resultou na remoção de 446.246 linhas.

O próximo passo foi a junção do *dataset* naturezas, que agora contém 22.095 registros, com o *dataset* vítimas, utilizando o método `.join()` e o atributo 'id_boletim' como chave. Como explicado na seção que apresenta a base de VCM construída nesta pesquisa, é possível ter mais de uma vítima para o mesmo boletim de ocorrência. Portanto, utilizando o tipo *inner join* para junção das bases naturezas e vítimas, esse procedimento gerou um *dataset* com 22.645 linhas e 21.884 'id_boletim' distintos.

Com esse novo *dataset* como tabela principal, foi feito novamente o *join*, agora com a base de autores e usando o tipo *left join*, para que todos os registros do *dataset* de naturezas e vítimas fossem mantidos e somente preenchidos quando houvesse correspondência na base de autores. Tal decisão se deu pelo fato de haver uma quantidade consideravelmente menor de registros da base de autores do que em todas as outras. Essa última junção gerou um novo *dataset*, agora com 22.669 linhas, o que indica que há crimes em que há mais de um autor. Durante o processo de junção dos arquivos que compõem a base paranaense, ficou evidente que todas as colunas

da base ocorrências também estavam presentes nos três outros arquivos, portanto esse arquivo acabou por ser desconsideração.

Tendo criado o conjunto de dados do Paraná que contém atributos que descrevem naturezas de crimes, vítimas e autores em um único *dataframe*, foram removidas as colunas que não tinham correspondência na base do Pará, utilizando o método `.drop()` que resultou em um *dataframe* com 29 colunas.

O próximo passo foi tratar as colunas `'data_fato'` e `'data_registro'`. Ambas eram do tipo *timestamp* e continham tanto a data quanto a hora do crime e do boletim de ocorrência. Para tanto, através do método `.getItem()` foi armazenado em uma coluna apenas o valor da primeira posição (0), que se refere à data, e na outra foi armazenado o valor da segunda posição (1) do campo *timestamp* de ambos os atributos, que se refere à hora. Analisando o tipo de atributos da base do Paraná com o método `.printSchema()`, percebeu-se que essa divisão tornou os campos de data tipados como *string*. Portanto, utilizando o método `.cast()` foi feita a tipagem dos atributos `'data_fato'`, `'data_registro'`, `'vit_dt_nasc'` e `'aut_dt_nasc'` para *date*.

Assim como na base do Pará, na base paranaense foi necessário aplicar um filtro no atributo `'ano_fato'`, para excluir os valores referentes ao ano de 2023, visto que não há esse ano na base paraense para comparação. O filtro de ano removeu do *dataset* 4.971 linhas. Em seguida, utilizando a função `lower()`, foi feita a substituição dos valores do atributo `'dia_semana'`, que eram apresentadas em letra maiúscula, para letra minúscula, em conformidade com os outros atributos e também com a base do Pará.

A base do Paraná também foi disponibilizada com atributos referentes à faixa horária de acontecimento do crime. Porém, assim como no *dataset* paraense, foi necessário fazer substituição de valores com base no atributo `'hora_fato'`, para que houvesse compatibilidade entre os valores de ambas as bases. Utilizando o método `.when()`, a hora de acontecimento do crime foi classificada em faixas horárias com valores como `'Manhã'`, `'Tarde'`, `'Noite'` e `'Madrugada'`, para a faixa horária a cada 6 horas, e `'20h – 22h'`, para a faixa horária a cada 2 horas.

Na base paranaense também foi feita a substituição dos valores já existentes dos atributos de idade pelo resultado da subtração entre a data do fato e a data de nascimento, visto que haviam erros de digitação. Então, com os novos valores para as idades, tanto da vítima quanto do autor, foi feita a reclassificação em faixas etárias. Ambas substituições foram feitas utilizando o método `.when()`. Também foram

encontrados valores inválidos para data de nascimento da vítima, como '1111-11-11', então as idades em que o ano de nascimento continha esse valor foram substituídas pela categoria 'Desconhecida'.

Finalizando o tratamento aplicado à base do estado do Paraná, utilizando o método `.withColumnRenamed()`, foram renomeadas algumas colunas para que o nome ficasse igual as colunas da base do Pará. Ainda, foi criada uma nova coluna para identificar o registros pertencentes ao estado, utilizando o método `.lit()`, acrescentando à todas as linhas o valor 'PR'. O último passo foi a ordenação das colunas, para que ficassem na mesma sequência das colunas da base do estado do Pará. O *dataset* final dos crimes de violência contra a mulher do estado do Paraná tem 17.698 registros e 33 atributos.

Tendo aplicado todos os tratamentos às bases de cada estado individualmente, exportou-se cada uma para um arquivo `.csv`, para posteriormente serem carregados em um novo *notebook* do Google Colab, a fim de iniciar os tratamentos da base unificada em um ambiente limpo de códigos. Tendo instalado o PySpark nesse novo Colab, importado as bibliotecas e iniciado a sessão Spark, foi possível carregar os arquivos `.csv` resultantes do tratamento das bases de cada estado individualmente.

Para unificar os dois *datasets* em um único foi utilizado o método `.union()`, que resultou em um conjunto de dados com 24.594 registros e 32 atributos. Após visualizar o *dataframe* resultante com o método `.show()`, identificou-se que a união dos arquivos fez com que os campos referentes à hora do crime e do registro do boletim de ocorrência fossem tipados como *timestamp*, onde a primeira posição assumiu a data do dia em que foi feito o procedimento. Portanto, utilizou-se o método `.getItem()` para manter apenas a posição da hora (1) de ambos os atributos.

Após manipulação mais aprofundada nos dados para a descrição estatística da base unificada, percebeu-se que havia duplicidade de alguns registros. Dessa maneira, através do método `.dropDuplicates()`, foi feita a remoção dos registros duplicados do *dataset*, que ficou com 24.534 linhas, a fim de evitar redundância na análise. Percebeu-se ainda que, para realizar o agrupamento de valores dos atributos em quantidade menor de categorias, seria necessário primeiramente lidar com os dados ausentes ou inconsistentes. Nesse sentido, foi usado o método `.fillna()` e o método `.replace()` para preencher e substituir com 'Desconhecido(a)' os campos que

são do tipo *string* onde os valores eram nulos ou contendo erros de digitação, respectivamente.

O agrupamento de valores de atributos mencionado anteriormente foi necessário tanto para a análise descritiva da base quanto para a aplicação do algoritmo de mineração. Tal procedimento foi necessário pelo fato de que havia atributos como 'local_ocorrenciã', 'meio_empregado' e 'grau_de_relacionamento' terem muitos valores possíveis, o que dificultaria a análise dos dados. Para tanto, foi realizada reclassificação desses atributos em menor quantidade de categorias. O atributo 'local_ocorrenciã' possuía 297 valores distintos, muitos deles contendo mais de um local na mesma categoria, como o valor 'residencia/rodovia estadual/sitio/veiculo em transito/via publica'. Nesse sentido, optou-se por selecionar apenas os trechos da *string* localizados antes do primeiro caractere '/', procedimento que resultou em 146 valores distintos.

Então, foram definidas 4 categorias para agrupar os locais de ocorrência de VCM, a saber: 'estabelecimento', que contém valores como 'restaurante', 'bar', 'estabelecimento de ensino'; 'residencia', que agrupa todos os valores onde há palavras como 'residencia', 'casa', 'moradia'; 'via publica', que contém valores como 'avenida', 'rua', 'via publica'; e 'outros', que agrupa os valores que não se encaixavam nas categorias anteriores.

O atributo 'meio_empregado' foi o que apresentou maior desafio para a reclassificação, visto que contava com 995 registros únicos, pois, assim como o atributo 'local_ocorrenciã', possuía muitos meios no mesmo valor, separados por '/'. Portanto, também foram selecionados apenas os valores antes do primeiro '/', resultando em 94 valores únicos, que foram reclassificados em 7 categorias diferentes, tais quais: 'ameaca', 'agressao', 'armas_em_geral', 'automovel', 'homem', 'sem_instrumento', 'via_digital' e 'outros'.

O próximo atributo que sofreu agrupamento de categorias foi o 'grau_de_relacionamento', que continha 57 valores distintos. Definiu-se então 5 categorias para descrever os diferentes valores encontrados para esse atributo, tais quais: 'familiar', 'pai/padrasto', 'parceiro_intimo', 'ex_parceiro_intimo' e 'outros'. O último atributo a passar por tratamento foi o 'registros', que representa o tipo de violência contra a mulher. Na base unificada havia 62 valores distintos para esse atributo, fato que compromete a análise se não observado. Nesse sentido, após avaliação de cada valor, definiu-se 9 categorias genéricas para reclassificação dos

valores. Esse procedimento possibilitou a identificação do atributo 'registros' como atributo meta para a classificação dos modelos de *machine learning* a serem utilizados na etapa de mineração dos dados. Os procedimentos de reclassificação de atributos descritos anteriormente foram possíveis através dos métodos `.when()`, `.isin()`, `.like()` e `.otherwise()`.

Todo o processo de pré-processamento descrito anteriormente resultou em um arquivo `.csv` contendo 24.534 registros e 33 atributos, o qual será analisado nas próximas seções. Tendo descrito anteriormente todos os passos executados para que fosse possível fazer a análise proposta nesta pesquisa, as duas próximas seções se ocupam em apresentar os encaminhamentos metodológicos que levaram ao processamento e o pós-processamento dos dados.

3.2.3 Processamento

A etapa de processamento de dados dessa pesquisa compreende os passos do processo KDD de análise exploratória dos dados, definição dos métodos de mineração, seleção do algoritmo de mineração e a efetiva mineração de dados.

A estatística descritiva nesta pesquisa se manifesta através de três tarefas principais: agrupamento de atributos, distribuição de contagem e apresentação das medidas de resumo. Todas as tarefas de análise exploratória de dados foram realizadas com o módulo `pyspark.sql`, do PySpark, no ambiente de desenvolvimento do Google Colab. O agrupamento de atributos teve como principais métodos o `.groupBy()`, `.count()`, `.orderBy()`, `.filter()` e `.select()`. Para a distribuição de contagem, os métodos `.pivot()` e `.show()` permitiram visualização de *dataframes* no *notebook* Colab, além de elaboração de gráficos usando as bibliotecas Matplotlib e Seaborn. Quanto à análise das medidas de resumo, os principais métodos utilizados foram `.agg()`, `.approxQuantile()` e `.freqItems()`; as funções utilizadas foram `mean` e `stddev`.

Após conhecer os dados com maior profundidade através da análise descritiva, foi necessário definir qual o método de mineração de dados mais apropriado para o objetivo do processo KDD proposto. Desse modo, optou-se pela utilização de tarefas preditivas, pois se ocupam em encontrar padrões para predição de comportamentos futuros sobre determinado problema. Para tanto, definiu-se pela aplicação de algoritmos de classificação, visto que grande parte dos atributos

presentes na base possui número discreto e finito de categorias. Foram escolhidos os algoritmos Random Forest e Multilayer Perceptron.

Random Forest (RF), proposto por Breiman (2001), é um algoritmo de aprendizado supervisionado que se propõe a resolver problemas de classificação e regressão. Trata-se de um método *ensemble*, pois utiliza a combinação de diferentes técnicas de predição a fim de potencializar os resultados. A principal ideia por trás do Random Forest é a construção de múltiplas árvores de decisão e a combinação de seus resultados para obter uma previsão mais robusta e precisa. Cada árvore é treinada com uma amostra aleatória do conjunto de dados e uma seleção aleatória de atributos (*features*). Essa abordagem introduz diversidade no conjunto de modelos, reduzindo assim o risco de *overfitting* e aumentando a generalização do modelo para novos dados. O voto da maioria das árvores no caso de classificação ou a média de suas previsões no caso de regressão é utilizado para determinar a saída final do Random Forest. Esse processo de "votação", ou combinação de resultados, ajuda a mitigar erros individuais das árvores, resultando em um modelo mais robusto e menos suscetível a variações nos dados de entrada.

O Multilayer Perceptron (MLP) é um modelo de rede neural artificial amplamente utilizado em aplicações de aprendizado profundo. Estruturado em camadas de neurônios, o MLP compreende uma camada de entrada que recebe os dados, uma ou mais camadas ocultas responsáveis por aprender representações complexas e uma camada de saída que gera as previsões. Cada neurônio realiza transformações não lineares nas entradas, possibilitando a captura de padrões e relações mais intrincadas nos dados. Durante o treinamento, os pesos das conexões entre neurônios são ajustados iterativamente por meio do algoritmo de retropropagação, minimizando uma função de perda. Essa arquitetura permite ao MLP aprender representações hierárquicas e realizar tarefas de classificação e regressão em diversos domínios.

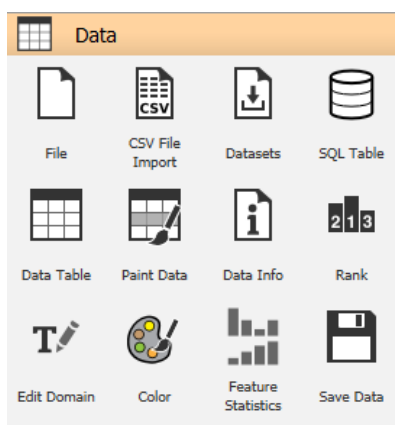
Com o conjunto de dados devidamente pré-processado e a metodologia, algoritmos e ferramenta de mineração de dados definidos, deu-se início ao processo de mineração utilizando o software Orange. Como mencionado anteriormente, o Orange destaca-se por sua interface intuitiva, onde o fluxo do processo de mineração é conduzido pelo inter-relacionamento de componentes, denominados na ferramenta como "widgets". Esses widgets, que desempenham papéis específicos, podem ser

facilmente inseridos e conectados no workflow, proporcionando uma abordagem visual e interativa para a construção e execução do processo de mineração de dados.

O Orange não apenas oferece uma variedade de algoritmos de mineração de dados prontos para uso, mas também permite a personalização e a combinação flexível desses algoritmos, proporcionando uma experiência robusta e adaptável para explorar *insights* e padrões nos dados. Essa abordagem orientada por widgets simplifica a criação de pipelines de mineração de dados, permitindo que os usuários se concentrem na análise dos resultados e na interpretação das descobertas. Tendo em vista que a autora não possui experiência prévia na ferramenta Orange, o aprendizado sobre a utilização se deu pela tentativa e erro de combinação de widgets, além de leitura da documentação do próprio software, que fornece explicação sobre os widgets e melhores práticas de uso para o *workflow*.

Para fins do processo de mineração de dados desta pesquisa, foram usados 7 widgets de 3 módulos diferentes, apresentados a seguir. O módulo 'Data' suporta atividades como leitura de arquivos nos mais diversos formatos, como .csv, .sql, .txt e .xlsx; visualização dos dados importados em formato de tabela; informações sobre a estrutura de dados, como quantidade de linhas e colunas, além de outras funcionalidades que permitem conhecer a base de dados. Dentre as opções possíveis, visualizadas na figura 9, foi selecionado o widget 'CSV File Import' para importação do arquivo pré-processado com PySpark. Visto que esse é um widget de início de *workflow*, de acordo com a documentação da ferramenta ele possui apenas conexões de saída, que podem ser ligadas a *widgets* como 'Data Table', 'Data Info' e 'Feature Statistics'.

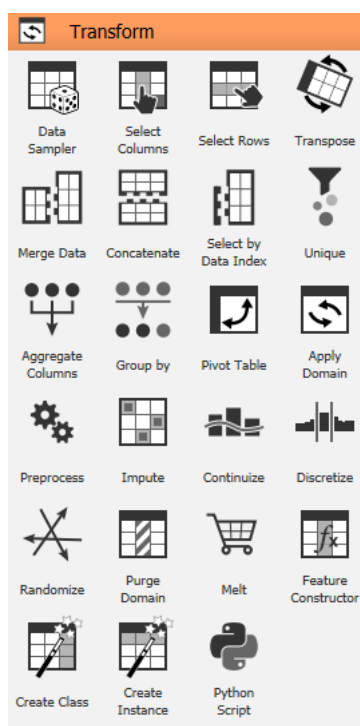
FIGURA 9 – WIDGETS DO MÓDULO 'DATA' DO ORANGE DATA MINING



FONTE: A autora (2023).

Após selecionar o arquivo dentro das configurações do widget 'CSV File Import', conectou-se ele ao 'Create Class', do módulo 'Transform'. Esse módulo fornece suporte a diversas atividades de pré-processamento de dados, como discretização e continuação de atributos, concatenação e junção de *datasets*, agregação de colunas, dentre outras, conforme ilustrado na figura 10. A utilização do widget 'Create Class' foi necessária para que fosse definido um atributo meta a ser considerado para a modelagem dos algoritmos de predição Random Forest e Multilayer Perceptron, visto que algoritmos de classificação precisam de um atributo que classifique os dados de acordo com a combinação dos atributos descritores. Esse widget tem tanto conexões de entrada quanto de saída; a entrada deve ser um *dataset* e a saída é um novo *dataset*, agora com a definição do atributo meta.

FIGURA 10 – WIDGETS DO MÓDULO 'TRANSFORM' DO ORANGE DATA MINING

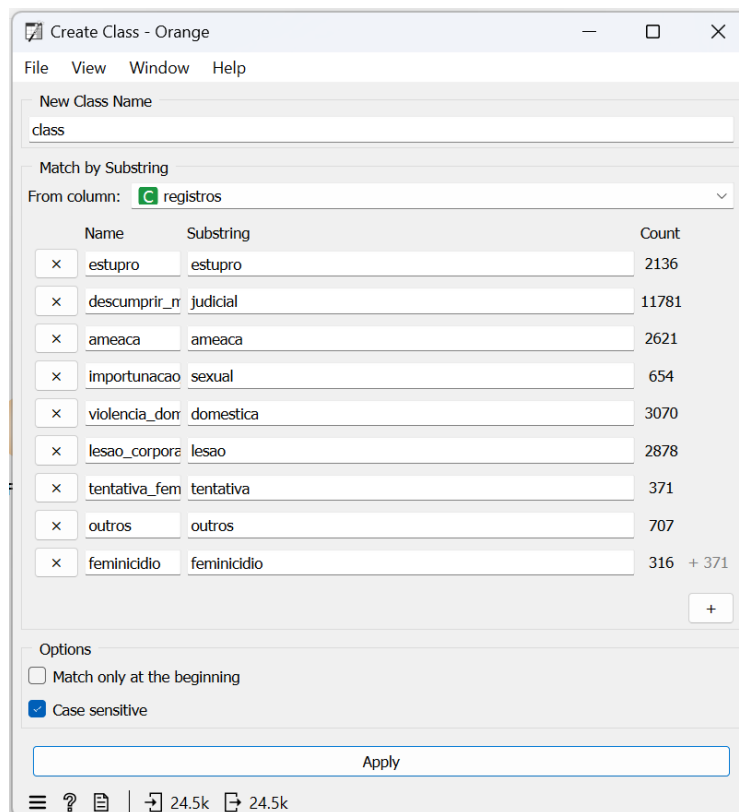


FONTE: A autora (2023).

Visto que na etapa de pré-processamento dos dados foi construído o atributo 'registros', que representa os diferentes tipos de violência contra a mulher presentes na base, a definição do atributo meta, tratado nos métodos e ferramentas de *machine learning* como 'class', se deu através da seleção da coluna 'registros' na tela de

configuração do widget 'Create Class'. Após selecionada a coluna, as categorias desejadas para o atributo meta foram obtidas através de combinação de *strings*, conforme detalhado na figura 11.

FIGURA 11 – TELA DE CONFIGURAÇÃO DO WIDGET 'CREATE CLASS'



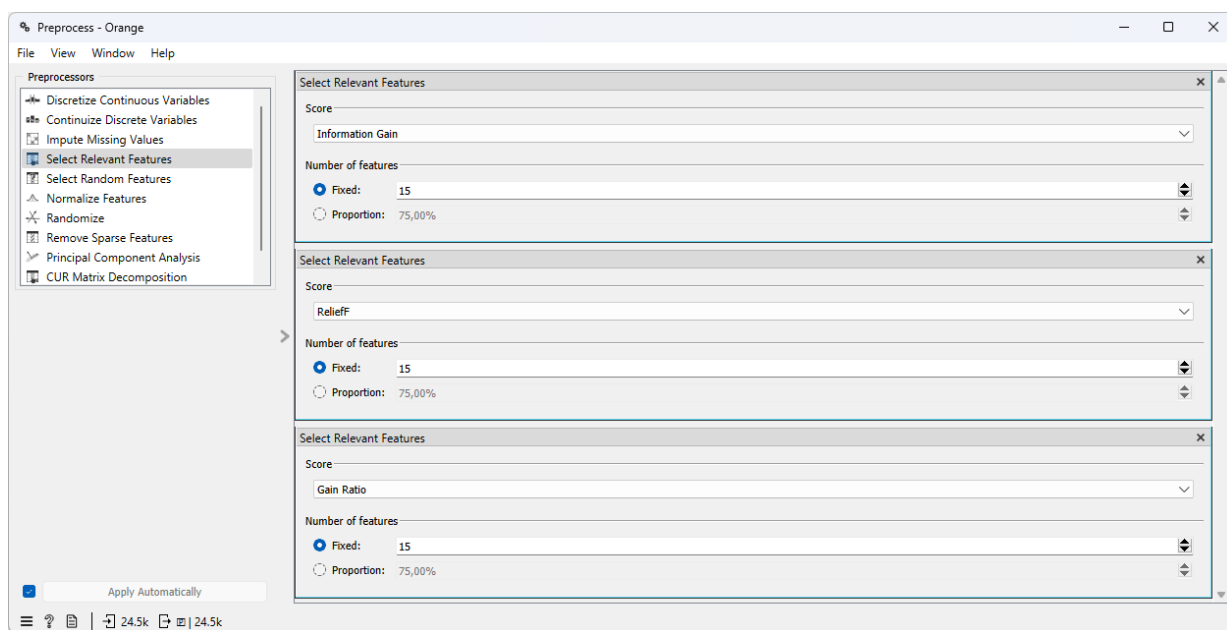
FONTE: A autora (2023).

De acordo com Costa (2022), 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. Desse modo, inseriu-se ao *workflow* o *widget* 'Preprocess', também do módulo 'Transform', que suporta várias técnicas de pré-processamento, como discretização de variáveis contínuas, normalização de atributos, imputação de valores ausentes e a seleção de *features* relevantes, sendo essa última a utilizada no *workflow*. O widget 'Preprocess' possui conexões de entrada, que devem ser conjuntos de dados, e também de saída, que são os métodos de pré-processamento e os dados após a manipulação.

Costa (2022) descreve a seleção de *features* como uma lógica apoiada na procura de um conjunto reduzido de atributos que melhore o desempenho dos algoritmos sem perder a qualidade dos resultados. Nesse sentido, realizou-se teste

de performance em 3 diferentes tipos de técnicas de redução de dimensionalidade de *features* na tela de configuração do widget ‘Preprocess’, representados na figura 12, sendo elas: Info Gain, Gain Ratio e ReliefF. Embora tenham sido apresentados todos as técnicas de filtragem de atributos em apenas uma tela, a execução do *workflow* foi feita utilizando uma técnica por vez.

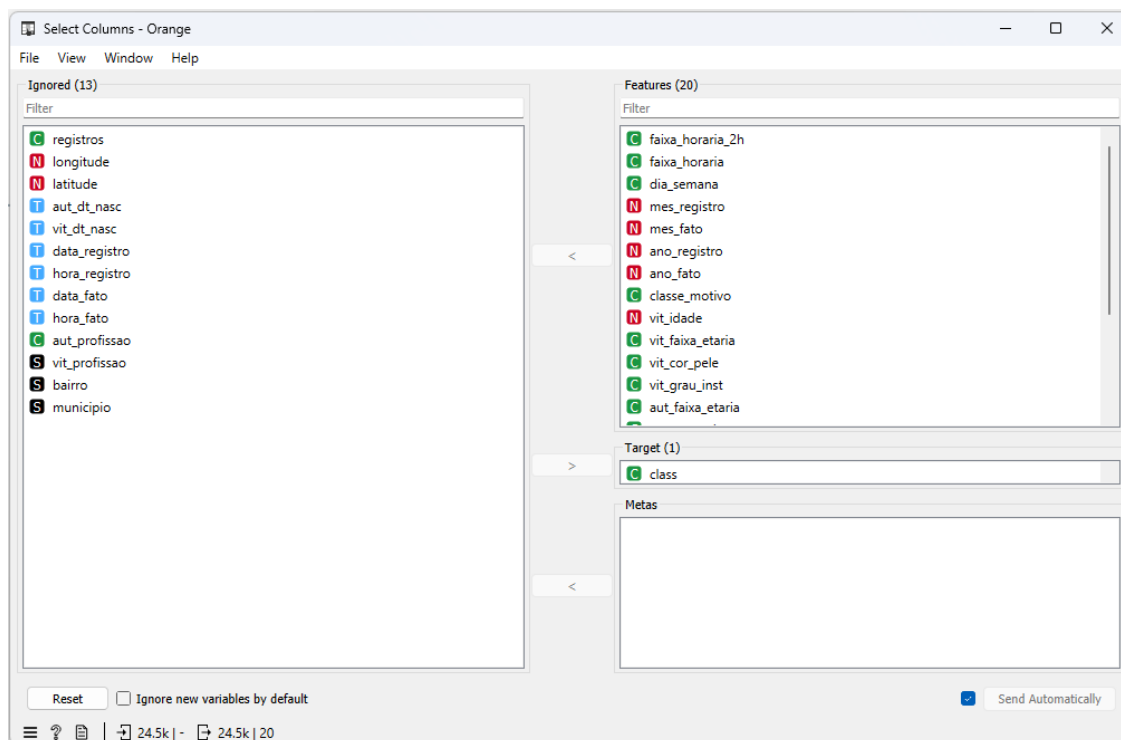
FIGURA 12 – TELA DE CONFIGURAÇÃO DO WIDGET ‘PREPROCESS’



FONTE: A autora (2023).

Ainda do módulo ‘Transform’, outro widget utilizado foi o ‘Select Columns’. Nesse componente é possível selecionar as colunas que serão efetivamente consideradas nas análises, assim como nos widgets anteriores, há conexões de entrada e de saída. Optou-se pela utilização desse componente pelo fato de haver alguns atributos que não apresentam relevância para a mineração de dados, como os de datas (do crime, do registro do boletim de ocorrência, de nascimento da vítima e do autor), de horas, os com elevado número de categorias, bem como os atributos ‘latitude’ e ‘longitude’. A figura 13 apresenta a seleção das colunas que serão utilizadas, localizadas à direita da tela de configuração do widget, bem como as colunas removidas, à esquerda, além da separação do atributo meta, identificado como ‘class’.

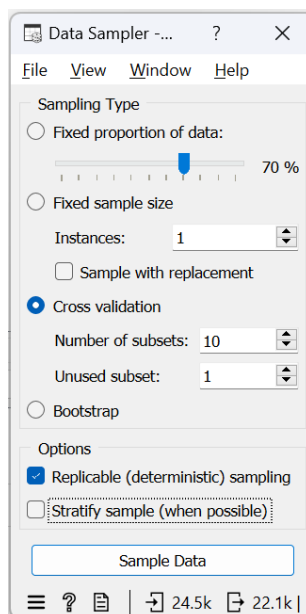
FIGURA 13 – TELA DE CONFIGURAÇÃO DO WIDGET ‘SELECT COLUMNS’



FONTE: A autora (2023).

O último widget utilizado do módulo ‘Transform’ foi o ‘Data Sampler’, que tem suporte à implementação de vários métodos de amostragem de dados. Possui conexões de entrada e de saída, recebendo conjuntos de dados e retornando as amostras selecionadas de acordo com as configurações de tipo de amostragem, respectivamente. Na tela de configuração do widget, ilustrada na figura 14, é possível definir entre quatro métodos de amostragem: proporção fixa de dados, que retorna uma porcentagem selecionada de dados (por exemplo, 70% de todos os dados); tamanho fixo de amostra, que retorna um número selecionado de instâncias com a opção de reposição de amostras; validação cruzada, que divide instâncias de dados em subconjuntos complementares e, por fim, o *bootstrap*, que infere a amostra a partir da estatística da população. Foram testados todos os métodos de amostragem a fim de verificar os resultados provenientes de cada um, para, posteriormente, definir o que apresentou as melhores métricas de avaliação para a construção do *workflow* final.

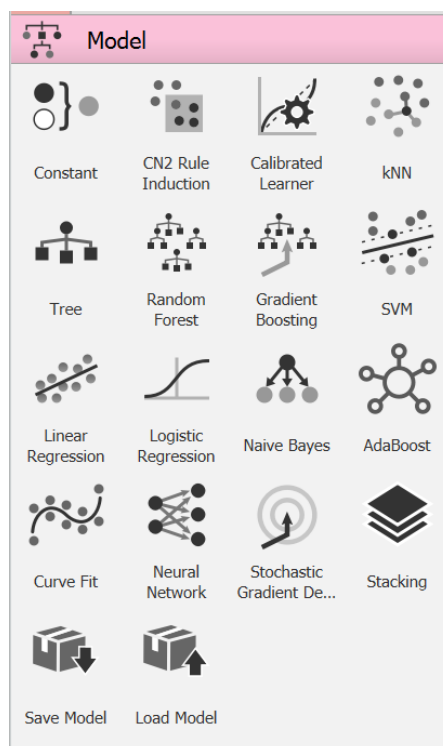
FIGURA 14 – TELA DE CONFIGURAÇÃO DO WIDGET ‘DATA SAMPLER’



FONTE: A autora (2023).

O próximo módulo utilizado na etapa de processamento foi o ‘Model’, que agrupa algumas técnicas e algoritmos de modelagem preditiva, como AdaBoost, Naive Bayes, SVM, kNN, dentre outros listados na figura 15. Para fins desta pesquisa, utilizou-se os widgets ‘Random Forest’ e ‘Neural Network’, baseado na rede neural MLP. Ambos possuem conexões de entrada, que podem ser widgets de fonte de dados ou de métodos de pré-processamento, e de saída, que são os algoritmos e modelos treinados. Tanto o widget ‘Random Forest’ quanto o ‘Neural Network’ possuem uma tela de configuração de parâmetros, que podem ser ajustados a fim de obter modelos mais robustos e resultados mais precisos. A configuração dos parâmetros para cada um dos widgets será apresentada a seguir.

FIGURA 15 – WIDGETS DO MÓDULO ‘MODEL’ DO ORANGE DATA MINING



FONTE: A autora (2023).

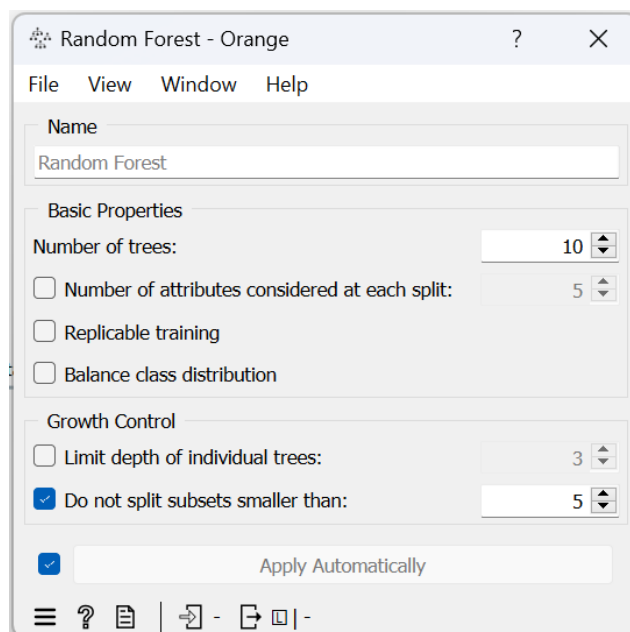
Nesse primeiro momento, são descritos os parâmetros de configuração do widget ‘Random Forest’. De acordo com a documentação oficial do Orange, esse widget pode ser aplicado tanto a tarefas de classificação quanto regressão; o que dita o comportamento do componente quanto a qual tarefa será executada são as conexões de entrada feitas, além do tipo de dado a ser predito. As configurações do widget ‘Random Forest’ são encontradas em dois principais grupos, a saber: propriedades básicas e controle de crescimento, além da opção de renomear o modelo, que por padrão recebe o nome do algoritmo.

Nas propriedades básicas podem ser definidas quantas árvores serão incluídas na floresta; o número de árvores consideradas em cada divisão, onde podem ser especificados quantos atributos serão aleatoriamente sorteados para consideração em cada nó; a definição de uma *seed* para geração da árvores, que permite a replicabilidade do treinamento e, por fim, a opção de balancear a distribuição do atributo classe, que, segundo a documentação da ferramenta, é inversamente proporcional ao peso de cada classe.

Nas configurações de controle de crescimento há dois parâmetros passíveis de configuração, sendo eles a profundidade máxima que cada árvore individual pode

ter, além da opção de definir um número mínimo de subconjunto a serem divididos. Após diversos testes com combinações de configuração diferentes, optou-se por utilizar as configurações padrão do widget, conforme identificado na figura 16.

FIGURA 16 – TELA DE CONFIGURAÇÃO DO WIDGET 'RANDOM FOREST'



FONTE: A autora (2023).

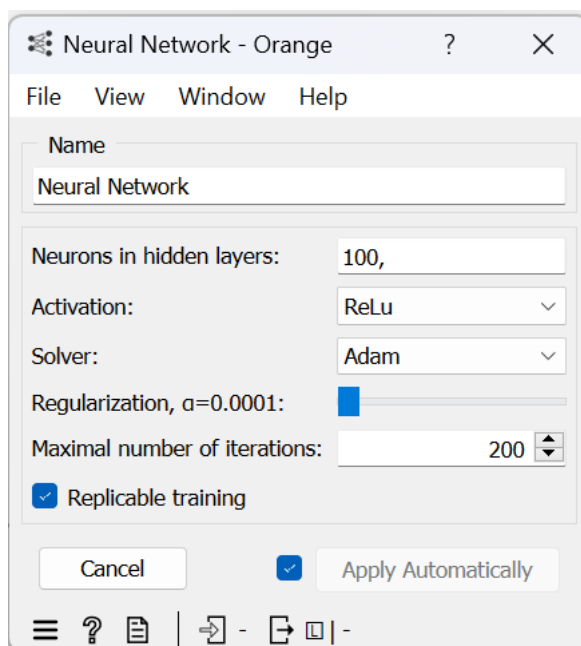
O último widget utilizado na etapa de processamento é o 'Neural Network'. Conforme a documentação do widget, o algoritmo de rede neural utilizado é o MLP da biblioteca Python Sci-kit Learn, que é capaz de aprender modelos lineares e não-lineares. Assim como no 'Random Forest', as conexões e entrada de dados é quem ditam o tipo de modelo treinado. A configuração do componente é ditada através de cinco parâmetros, além da opção de renomear o modelo, que por padrão é nomeado como 'Neural Network'.

O primeiro parâmetro configurável é a quantidade de neurônios, também chamados de perceptrons, presentes em cada camada intermediária (ou camada oculta) da rede neural. O próximo parâmetro é a função de ativação da camada oculta, que conta com quatro opções de funções, a saber: 'Identity', indicada pela documentação do widget para implementação linear; 'Logistic', uma função sigmóide para aplicação não-linear; 'Tahn', uma função sigmóide escalonada, e a função não-linear 'ReLU'. O terceiro parâmetro passível de configuração é o solucionador para otimização de peso, que possui as opções 'L-BFGS-B', 'SGD' e 'Adam'. Ainda, é

possível configurar o parâmetro de regularização do modelo, também conhecido como pênalti, que visa mitigar o *overfitting*. O último parâmetro é o número máximo de iterações que o algoritmo executará.

Assim como no widget 'Random Forest', após testes com diferentes parâmetros, definiu-se que a configuração padrão do componente 'Neural Network' se saiu melhor na modelagem do algoritmo, conforme descrito na figura 17.

FIGURA 17 – TELA DE CONFIGURAÇÃO DO WIDGET 'NEURAL NETWORK'



FONTE: A autora (2023).

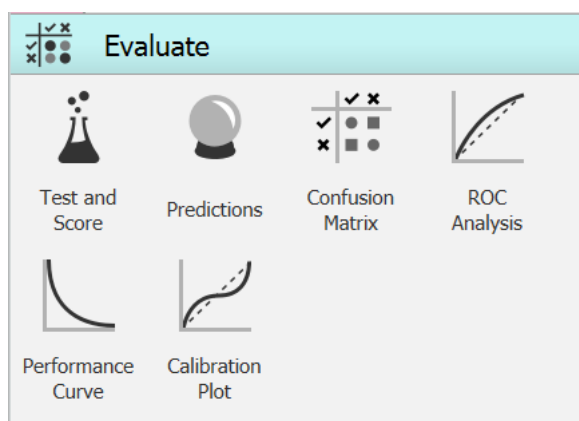
Descritos os procedimentos realizados na etapa de processamento no *software* Orange, a próxima seção apresenta os métodos utilizados para a avaliação dos resultados dos modelos de aprendizado de máquina desenvolvidos.

3.2.4 Pós-processamento

A última etapa metodológica do processo KDD desenvolvido nesta pesquisa cobre as tarefas inerentes à interpretação dos resultados obtidos através da modelagem da etapa de processamento. Para tanto, foram utilizados 3 widgets do módulo 'Evaluate' do *software* Orange Data Mining. Este módulo é especializado na avaliação da performance dos modelos inseridos no fluxo de trabalho, disponibilizando widgets que fornecem os resultados de diversas métricas de avaliação. Essas

métricas não apenas revelam as previsões individuais de cada modelo, mas também proporcionam visualizações elucidativas, como gráficos de regressão logística e linear. Complementarmente, é possível gerar matrizes de confusão, oferecendo uma representação concisa e informativa do desempenho dos modelos em questão. Este conjunto de ferramentas revelou-se fundamental para uma análise abrangente e aprofundada da eficácia dos modelos desenvolvidos. Dentre os widgets disponíveis, apresentados na figura 18, foram usados o 'Test and Score', 'Predictions' e 'Confusion Matrix'.

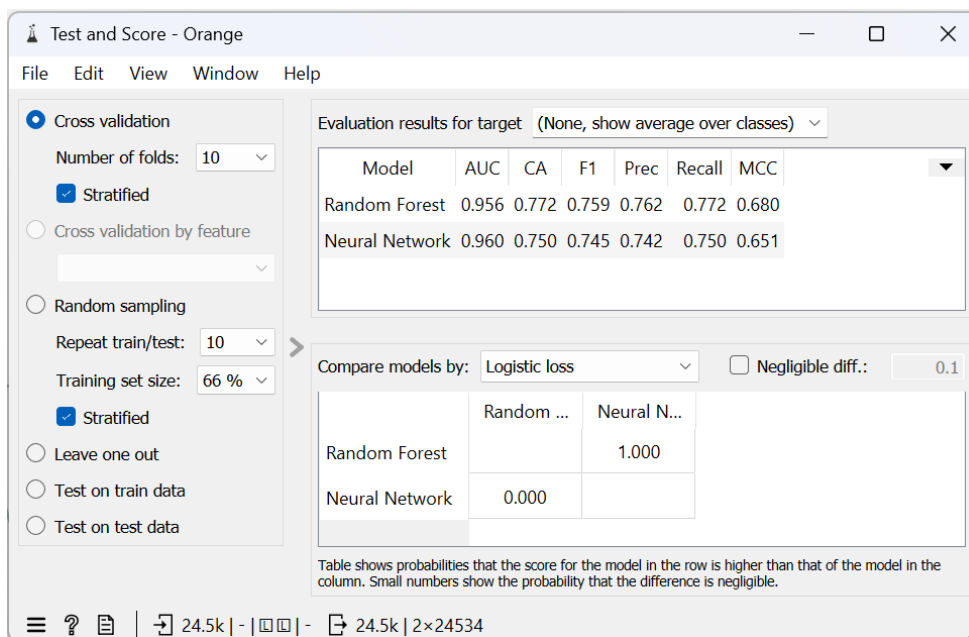
FIGURA 18 – WIDGETS DO MÓDULO 'EVALUATE' DO ORANGE DATA MINING



FONTE: A autora (2023).

O widget 'Test and Score' possui conexões de entrada e saída, com a entrada disponível para receber *datasets*, dados para teste e os algoritmos de aprendizado, e a saída retornando os resultados da avaliação dos testes de cada algoritmo utilizado no *workflow*. De acordo com a documentação da ferramenta, o widget se ocupa com duas tarefas: primeiro, ele mostra uma tabela com diferentes medidas de desempenho do classificador, como a precisão da classificação e a área sob a curva, para então gerar os resultados da avaliação, que podem ser usados por outros widgets para analisar o desempenho dos classificadores, como a Análise ROC ou a Matriz de Confusão. Nesse sentido, o componente oferece suporte a vários métodos de amostragem para testes, como a validação cruzada, validação cruzada por *feature* e amostragem aleatória, conforme descrito na figura 19, os quais foram todos testados a fim de definir o que apresentaria melhores resultados.

FIGURA 19 – TELA DE CONFIGURAÇÃO DO WIDGET ‘TEST AND SCORE’



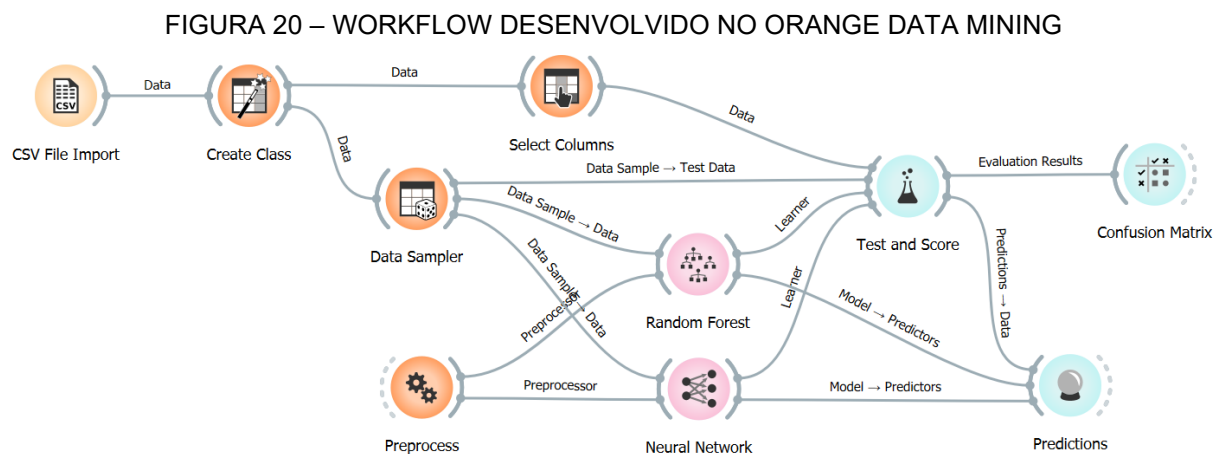
FONTE: A autora (2023).

Outro widget do módulo ‘Evaluate’ utilizado foi o ‘Predictions’. Como sugere o nome, esse componente apresenta as predições feitas para cada modelo e possui tanto conexões de entrada, para fontes de dados e modelos preditivos, quanto de saída, onde fornece a estrutura dos dados de entrada com as predições adicionadas e resultados de testes de algoritmos de classificação. Para a aplicação na presente pesquisa, as configurações disponíveis para o widget ‘Predictions’ se referem apenas a maneira de apresentação das predições, onde é possível definir se deseja-se visualizar as probabilidades de predição para cada valor do atributo meta ou a média de todas as categorias.

O último widget de avaliação dos resultados empregado foi o ‘Confusion Matrix’. Matrizes de confusão são ferramentas analíticas amplamente utilizadas para avaliar o desempenho de modelos de aprendizado supervisionado, onde são comparadas as predições feitas pelo modelo com os resultados reais presentes no *dataset*. Na conexão de entrada, esse componente espera resultados da avaliação dos algoritmos de classificação fornecidos pelo widget ‘Test and Score’ e como saída entrega a visualização da matriz de confusão, que pode ser configurada para apresentar o número de instâncias, a proporção (porcentagem) de predições, a proporção dos resultados reais ou a soma das probabilidades. Para apresentação dos

resultados da predição na matriz de confusão, optou-se por visualizar a proporção de predição do modelo.

Após a descrição detalhada de todos os widgets utilizados no processamento e pós-processamento de dados desta pesquisa, apresenta-se na figura 20 o *workflow* completo desenvolvido no *software* Orange Data Mining.



FONTE: A autora (2023).

Apresentado o trajeto metodológico percorrido durante o desenvolvimento desta pesquisa, a próxima seção se ocupa em apresentar os resultados obtidos através da aplicação do processo KDD.

4 RESULTADOS E DISCUSSÃO

Nesta seção são evidenciados os resultados obtidos através da aplicação das etapas de processamento e pós-processamento do processo KDD na base unificada, que contém os registros dos estados do Pará e Paraná após a aplicação de todos os filtros que caracterizam os dados como ocorrências de violência contra a mulher.

4.1 ESTATÍSTICA DESCRITIVA

De acordo com Castro e Ferrari (2016), a análise descritiva dos dados é o princípio de qualquer análise quantitativa de dados, usada para descrever, simplificar ou sumarizar as principais características de uma base de dados. Como parte da análise dos dados sobre VCM desta pesquisa, foi feita a estatística descritiva através de identificação de medidas de tendência central e de variabilidade, além de contagem agrupada de atributos apresentada em visualizações gráficas das distribuições dos dados e elaboração de tabelas de contagem.

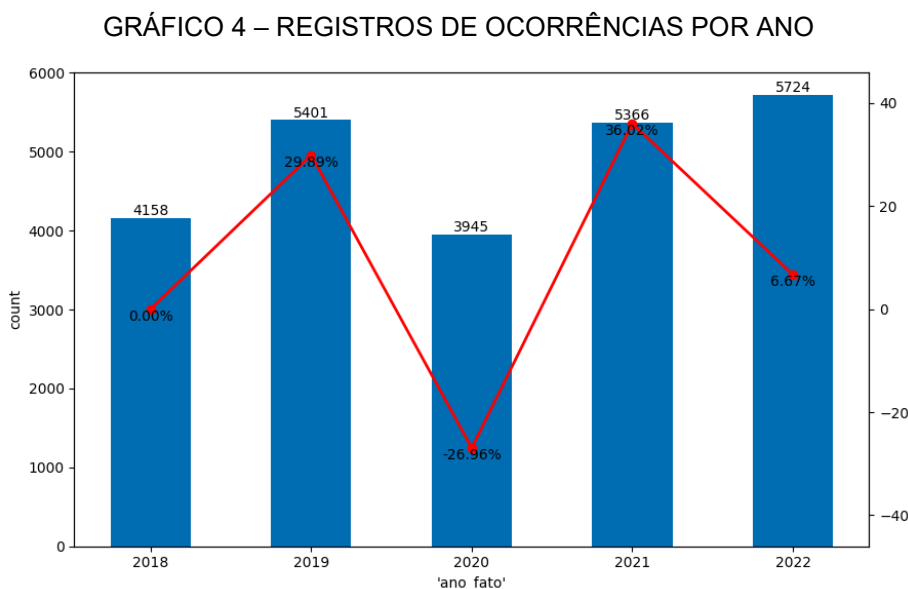
Para Monteiro e Caballero (2021), um critério de extrema relevância para a análise de dados criminais é a definição do período de apuração dos números, pois é possível contabilizá-los pela data de ocorrência do fato ou pela data de registro no sistema de informação da polícia. Para os fins desta pesquisa, optou-se por utilizar a data e hora do fato como base para a contabilização das ocorrências, visto que apenas em 97 casos houve diferença entre a data do fato e a data de registro, sendo que, destes, a grande maioria é de apenas 1 dia de diferença, com 73 ocorrências.

Através da distribuição dos dados em relação ao atributo 'ano_fato', evidenciada pelo gráfico 4, pode-se observar que o ano com maior incidência de violência contra a mulher foi 2022, com 5.724 registros, e menor incidência foi 2020, com 3.945 registros. A incidência média é de aproximadamente 4.907 casos por ano, com desvio padrão de 807,66 e coeficiente de variação de 16,46%, indicando variação considerável em relação à média.

Ao analisar o gráfico 4, também é possível identificar a variação percentual de incidência em relação ao ano anterior, representada pela linha vermelha, onde cabe chamar a atenção para a variação negativa de aproximadamente -30% entre 2019 e 2020. Monteiro e Caballero (2021, p. 145) apontam que é preciso ter muito cuidado ao analisar indicadores desse fenômeno, pois não é possível saber, “a priori, se de

fato esta violência está ocorrendo mais, ou se só estão sendo mais registradas devido à maior conscientização da população e o maior acesso às delegacias”. Segundo os autores, todas as variações abruptas em um indicador devem ser analisadas com bastante cautela, pois há diversos fatores que podem se correlacionar, mas não necessariamente terem relação de causa e efeito.

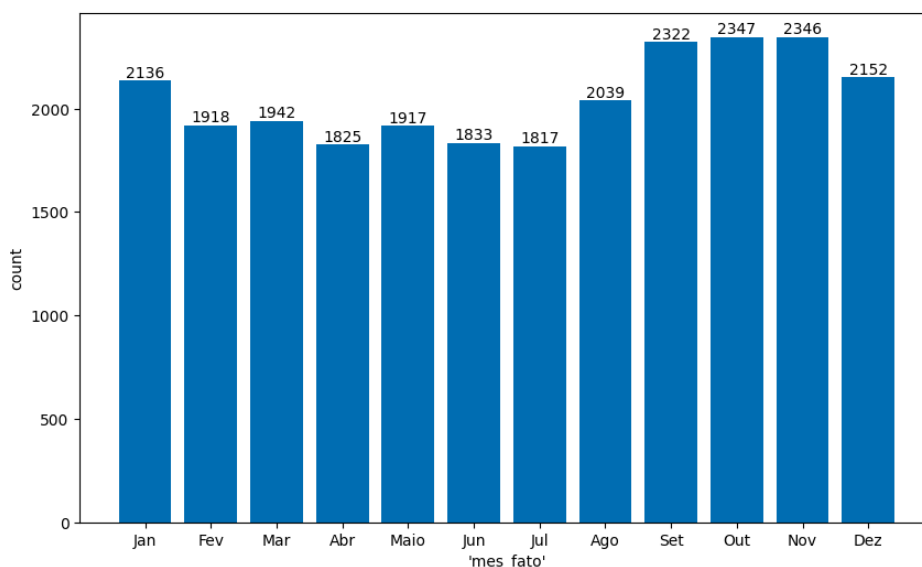
Entretanto, uma provável explicação para a brusca queda pode ser a subnotificação causada pelo isolamento social experienciado devido à pandemia do COVID-19, visto que o ano seguinte, quando já havia menores índices de isolamento, apresentou variação positiva acentuada em relação a 2020, com aumento de aproximadamente 36%.



FONTE: A autora (2023).

Aprofundando a análise dos registros em relação à dimensão de temporalidade, observa-se maior incidência de ocorrências de violência contra a mulher no segundo semestre do ano, sendo o mês de outubro o com a maior contagem de registros, contando com 2.347 casos seguido pelos meses de novembro e setembro, com 2.346 e 2.322 casos, respectivamente, conforme o gráfico 5. A incidência média mensal é de aproximadamente 2.044 ocorrências, com desvio padrão de 205,77 e coeficiente de variação de 10,06%, sugerindo variação moderada da ocorrência média mensal, sendo um fator positivo para previsibilidade de padrões.

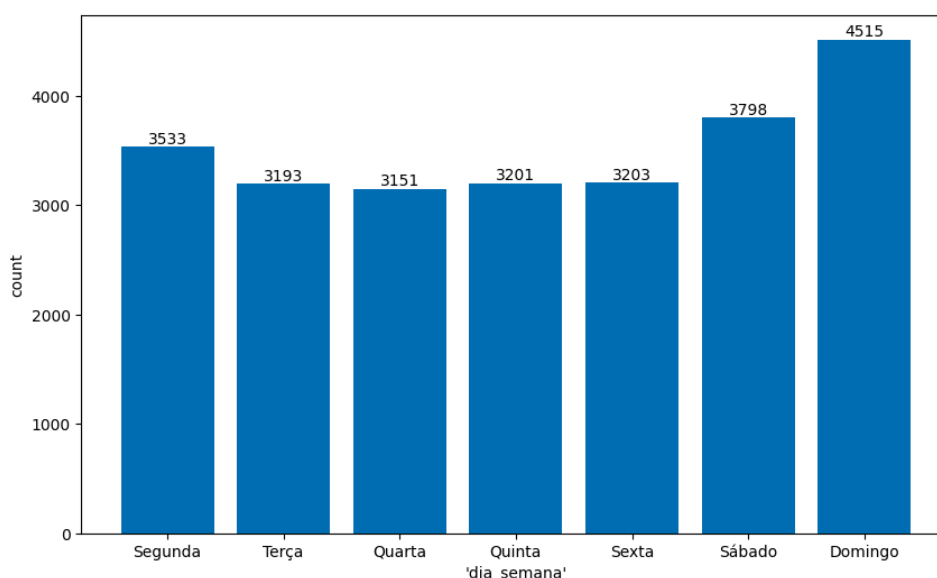
GRÁFICO 5 – REGISTROS DE OCORRÊNCIAS POR MÊS



FONTE: A autora (2023).

Ao analisar os registros das ocorrências por semana, nota-se que o fim de semana concentra grande parte dos casos semanais, com 3.798 registros para o sábado e 4.515 para o domingo, e o meio da semana é quando ocorrem menos casos, com 3.151 registros para a quarta-feira, segundo descrito no gráfico 6. A incidência média semanal é de aproximadamente 3.504 casos, com desvio padrão de 499,87 e coeficiente de variação de 14,26%.

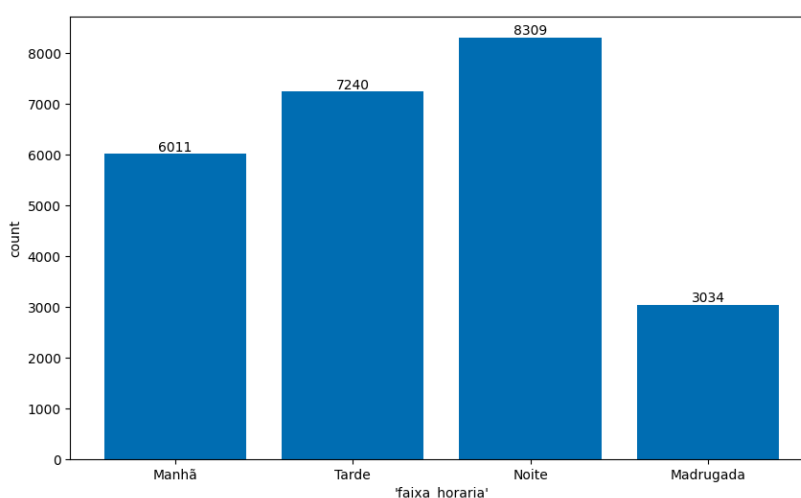
GRÁFICO 6 – REGISTROS DE OCORRÊNCIAS POR SEMANA



FONTE: A autora (2023).

Mediante a análise dos registros dos dados em relação aos distintos períodos do dia, procedeu-se à investigação do atributo 'faixa_horaria', construído por meio da categorização do atributo 'hora_fato' em quatro faixas horárias predefinidas, a saber: manhã (06h às 12h), tarde (12h às 18h), noite (18h às 00h) e madrugada (00h às 06h). Analisando o gráfico 7, é possível perceber que o horário em que há maior ocorrência de crimes contra a mulher é das 18 horas da tarde até a meia noite e o horário em que ocorrem menos crime é da meia noite até as 06 horas da manhã.

GRÁFICO 7 – REGISTROS DE OCORRÊNCIAS POR PERÍODO DO DIA



FONTE: A autora (2023).

Para concluir a análise temporal da base de ocorrências de violência contra a mulher, buscando compreender com maior clareza o horário em que mais ocorre esse tipo de crime, investigou-se os registros em relação ao atributo 'faixa_horaria_2h', também construído por meio da categorização do atributo 'hora_fato', agora agrupando os dados a cada 2h. Conforme descrito na tabela 1, o horário com maior registros de ocorrências é entre 18h e 20h da noite, com 3.024 registros, seguido por 20h às 22h, com 2.902 registros; a menor contagem observada é entre as 02h e 04h da madrugada, com 905 registros.

TABELA 1 – REGISTROS DE OCORRÊNCIAS POR FAIXA HORÁRIA

Faixa 2h	Registros
00h - 02h	1583
02h - 04h	905
04h - 06h	546

06h - 08h	1003
08h - 10h	2192
10h - 12h	2816
12h - 14h	2371
14h - 16h	2443
16h - 18h	2426
18h - 20h	3024
20h - 22h	2902
22h - 00h	2383

FONTE: A autora (2023).

Partindo para a análise de características demográficas, nesse primeiro momento descrevendo os dados em relação à vítima, analisou-se o atributo 'vit_idade' obtendo a idade média de aproximadamente 36 anos, com desvio padrão de 12,86 e coeficiente de variação de 35,93%. Através da análise das medidas de resumo do atributo 'vit_idade' para cada ano, detalhadas na tabela 2, é possível perceber que as idades das vítimas são altamente dispersas em relação à média, pois em todos os anos o coeficiente de variação é maior que 30%, indicando distribuição heterogênea dos dados.

TABELA 2 – MEDIDAS DE RESUMO DO ATRIBUTO IDADE POR ANO

Ano	Média	Desvio Padrão	Coeficiente de Variação (%)	1º Quartil	3º Quartil
2018	33,38	11,88	35,14	24	40
2019	35,21	12,67	34,75	25	42
2020	36,12	12,79	35,93	26	43
2021	36,11	13,02	37,57	26	44
2022	37,55	13,30	38,40	27	45

FONTE: A autora (2023).

A base de dados contém ainda um atributo que classifica as vítimas em relação à faixa etária que pertence, cuja classificação e contagem são descritas na tabela 3. A faixa etária com menos registros é a criança, com 26 ocorrências, ao passo que a faixa etária jovem I é a que agrupa a maior quantidade de registros, contando 4.580. Essa última observação põe em xeque a utilização somente da média de 36 anos informada anteriormente como medida de resumo da idade das vítimas, visto que a maior concentração de dados se refere a mulheres de 18 a 24 anos. Tal fato

pode ser explicado pela presença de valores extremos de idades, visto que a maior idade observada é 105 anos, referente a 1 registro, e a menor 0 anos, referente a 3 registros.

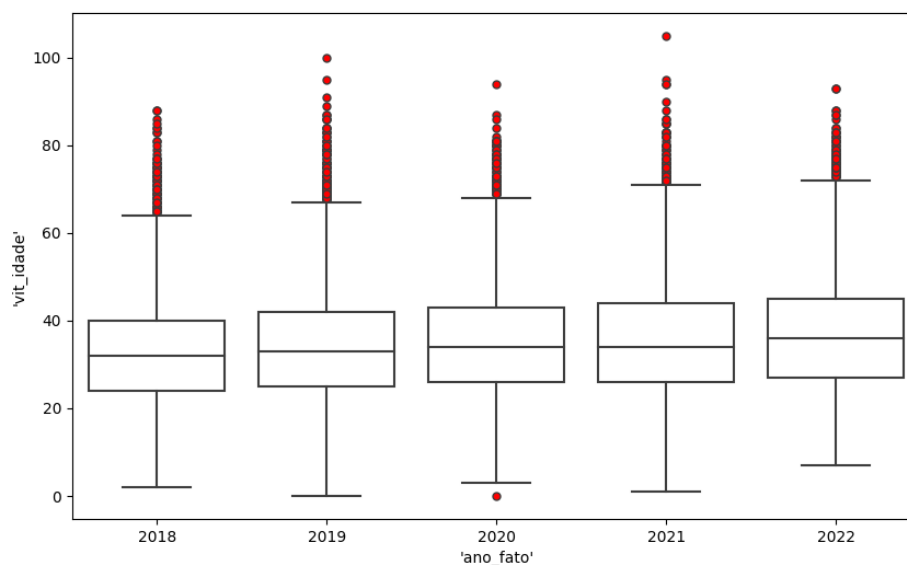
TABELA 3 – CLASSIFICAÇÃO E REGISTROS DO ATRIBUTO FAIXA ETÁRIA DA VÍTIMA

Faixa Etária	Idades	Registros
Criança	De 0 a 11 anos	26
Adolescente	De 12 a 17 anos	229
Jovem I	De 18 a 24 anos	4.580
Jovem II	De 25 a 29 anos	3.977
Adulto I	De 30 a 34 anos	3.863
Adulto II	De 35 a 39 anos	3.491
Adulto III	De 40 a 44 anos	2.723
Meia idade I	De 45 a 49 anos	1.909
Meia idade II	De 50 a 64 anos	2.526
Idosa	Mais de 65 anos	838
Desconhecida	Agrupamento de valores nulos ou erros de digitação	372

FONTE: A autora (2023).

Com o intuito de aprofundar a análise sobre a idade das vítimas proposta nas tabelas 2 e 3, o gráfico 8 ilustra a dispersão e amplitude dos dados em relação ao atributo 'vit_idade' agrupados por ano, evidenciando os valores que se distanciam muito da mediana, os chamados *outliers*, representados em vermelho. A presença dos *outliers* é um fator de forte influência na definição da média de 36 anos, uma vez que 50,6% das ocorrências são referentes a mulheres que possuem idade entre 18 a 34 anos, correspondente às faixas etárias jovem I, jovem II e adulto I.

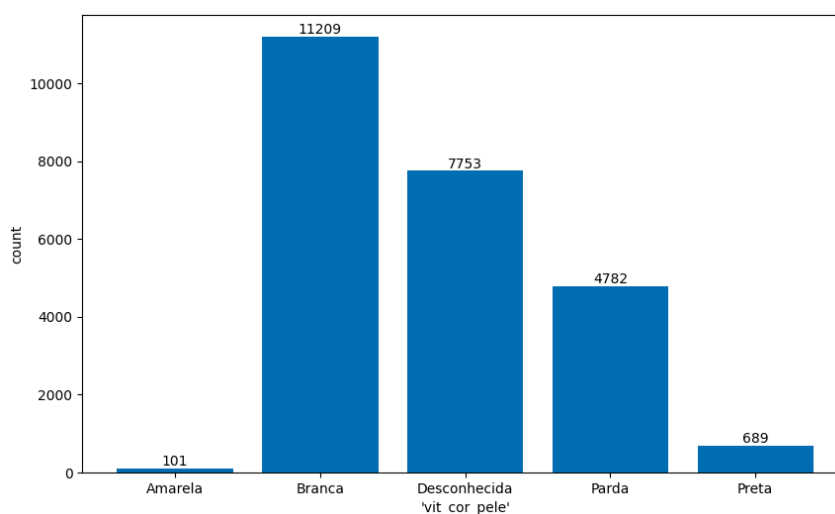
GRÁFICO 8 – DISPERSÃO DE IDADES DAS VÍTIMAS AGRUPADAS POR ANO



FONTE: A autora (2023).

Outra característica presente nos dados analisados é a cor da pele das vítimas, representada pelo atributo "vit_cor_pele". Ao examinar o registro das ocorrências com base na cor da pele das vítimas, conforme evidenciado no gráfico 9, torna-se notável que uma parcela substancial dos registros desse atributo apresenta valores nulos ou preenchimento inválido, correspondendo a 31,6% do total de ocorrências, categorizados sob a classificação "Desconhecida".

GRÁFICO 9 – REGISTROS DE OCORRÊNCIAS POR COR DA PELE DA VÍTIMA



FONTE: A autora (2023).

A expressiva ausência de dados válidos tem repercussão significativa na capacidade de conduzir uma análise precisa do contexto racial das vítimas, um fator que, por sua vez, afeta a formulação de políticas públicas voltadas para a promoção da igualdade na sociedade. Esse impacto é substanciado pela necessidade de as autoridades legisladoras adotarem abordagens diferentes para atender às demandas de mulheres pertencentes a etnias diferentes, uma vez que a cor da pele é capaz de gerar contextos sociais diversos. Isso se torna essencial para assegurar principalmente o direito de mulheres negras à segurança e integridade, visto que, conforme relatado no 17º Anuário Brasileiro de Segurança Pública, 61,1% dos casos de feminicídio e 56,8% dos estupros ocorridos em 2022 afetaram a população negra (FBSP, 2023).

A próxima característica explorada no conjunto de dados é o grau de instrução da vítima, cujo número de registros é apresentado no quadro 4. A escolaridade com maior número de registros válidos é o ensino médio completo, com 4.025 ocorrências, seguida pelo ensino fundamental incompleto, com 3.758 ocorrências. A categoria menos expressiva é 'não alfabetizada', com 263 registros. Assim como na análise do atributo 'vit_cor_pele', é possível perceber que grande parte das ocorrências não tem valor válido para o grau de instrução da vítima, contabilizando 43,8% das ocorrências, categorizadas como "Desconhecido".

Conhecer com profundidade o nível educacional de determinada população permite entender seu acesso a recursos, conscientização sobre direitos e riscos de vitimização a fim de direcionar ações públicas voltadas a grupos específicos. Nesse cenário, onde nota-se a educação frequentemente relacionada a desafios sociais e econômicos, a ausência de informações sobre a escolaridade pode prejudicar a capacidade de abordar questões de desigualdade educacional e gerar limitação na formulação de políticas públicas.

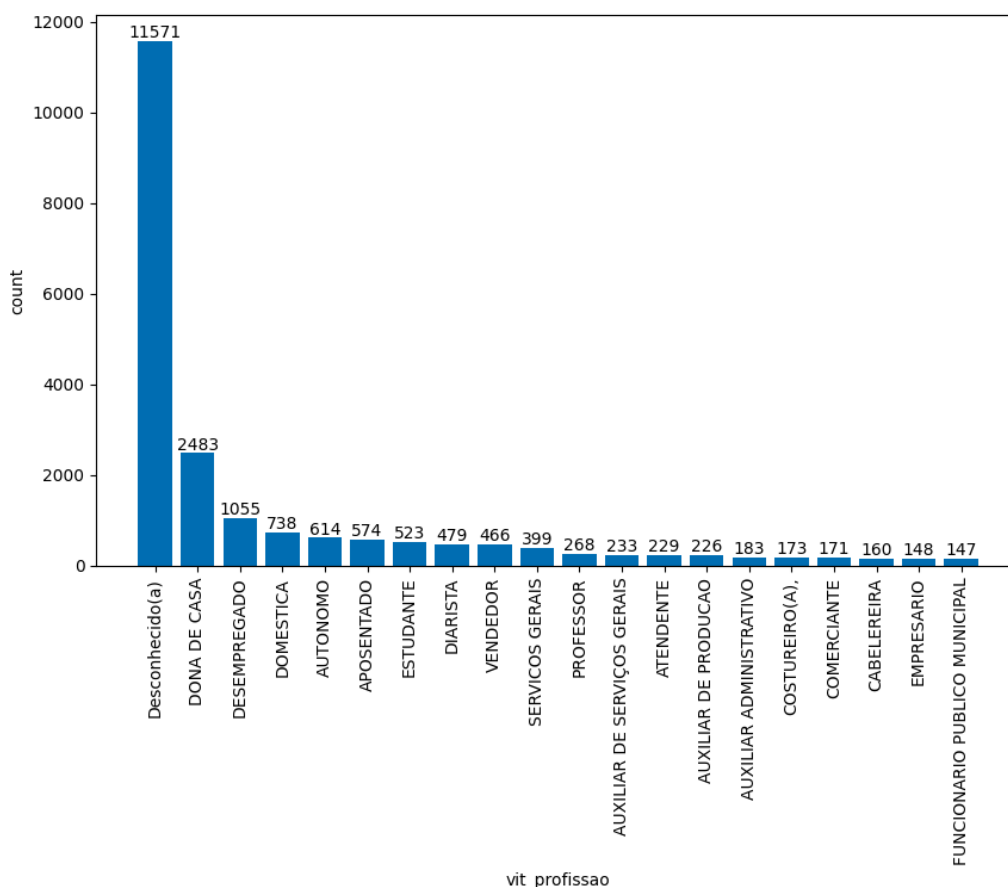
QUADRO 4 – REGISTROS DE OCORRÊNCIAS POR GRAU DE INSTRUÇÃO DA VÍTIMA

Grau de instrução da vítima	Registros
Desconhecido	10.759
Ensino Fundamental Completo	1.398
Ensino Fundamental Incompleto	3.758
Ensino Médio Completo	4.025
Ensino Médio Incompleto	1.640
Ensino Superior Completo	1.743
Ensino Superior Incompleto	948
Não alfabetizada	263

FONTE: A autora (2023).

Quanto aos atributos referentes à vítima, a última característica investigada nesta pesquisa é a sua profissão. O gráfico 10 mostra apenas as 20 categorias com maior número de ocorrências, pois há 295 valores distintos para o atributo 'vit_profissao', montante que impede apresentação de todas as profissões registradas no *dataset*.

GRÁFICO 10 – REGISTROS DE OCORRÊNCIAS POR PROFISSÃO DA VÍTIMA

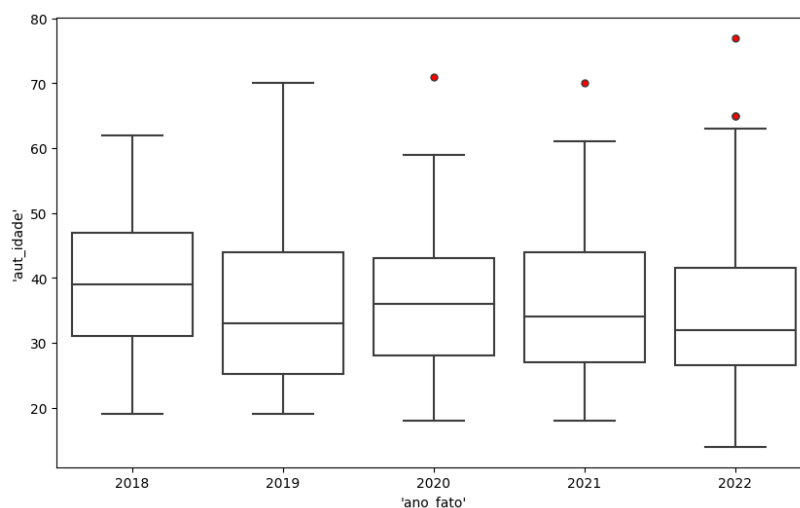


FONTE: A autora (2023).

Passando a descrever os dados quanto às características demográficas do autor do crime, cabe informar primeiramente que a quantidade de dados não válidos para os atributos relacionados ao autor é ainda maior do que àqueles observados para as vítimas. Para os atributos descritos a seguir, há menos de 10% de dados válidos relacionados ao autor, constatação que merece atenção, pois dificulta a identificação de padrões nos dados e descoberta de conhecimento do contexto social vivido pelos homens que cometem esse tipo de crime. Nesse sentido, optou-se por apresentar os dados referentes ao autor predominantemente em quadros, visto que a visualização de gráficos de agrupamentos de atributos fica prejudicada quando uma única categoria agrupa a maioria dos dados, identificados como 'Desconhecido(a)', categoria definida para quando o registro era nulo ou continha erro de entrada.

Dito isso, observando os dados em relação ao atributo 'aut_idade', verificou-se que, dos 24.534 registros do *dataset*, 24.051 são nulos, proporção que prejudica o conhecimento abrangente das características etárias dos autores de VCM. De todo modo, contabilizando os dados válidos, obteve-se a idade média de aproximadamente 36 anos, com desvio padrão de 11,78 e coeficiente de variação de 32,84%. A maior idade observada é 77 anos e a menor 14 anos, ambas contabilizando apenas 1 registro. O gráfico 11 apresenta a dispersão e amplitude dos dados em relação ao atributo 'aut_idade' agrupados por ano, evidenciando a presença de poucos *outliers*, representados em vermelho.

GRÁFICO 11 – DISPERSÃO DE IDADES DOS AUTORES AGRUPADAS POR ANO



FONTE: A autora (2023).

O próximo atributo examinado é a cor da pele dos autores, representada pelo atributo "aut_cor_pele". Pelo número de registros descrito no quadro 5, observa-se que a categoria 'branca' é a que possui o maior número de registros válidos, contabilizando 265, sendo os menores registros os referentes às categorias 'amarela' e 'indígena', com 1 registro cada.

QUADRO 5 – REGISTROS DE OCORRÊNCIAS POR COR DA PELE DO AUTOR

Cor de pele do autor	Registros
Amarela	1
Branca	265
Desconhecido(a)	24.113
Indígena	1
Parda	131
Preta	23

FONTE: A autora (2023).

O quadro 6 descreve os registros dos dados em relação ao grau de instrução do autor, presentes no atributo 'aut_grau_inst'. Ensino fundamental incompleto é a categoria que agrupa a maioria dos dados válidos, sendo 76; assim como no caso do grau de instrução da vítima, 'não alfabetizado' é a categoria com menor quantidade de ocorrências.

QUADRO 6 – REGISTROS DE OCORRÊNCIAS POR GRAU DE INSTRUÇÃO DO AUTOR

Grau de instrução do autor	Registros
Desconhecido(a)	24.346
Ensino Fundamental Completo	18
Ensino Fundamental Incompleto	76
Ensino Médio Completo	41
Ensino Médio Incompleto	15
Ensino Superior Completo	19
Ensino Superior Incompleto	16
Não alfabetizado	3

FONTE: A autora (2023).

Continuando a descrição das características do autor, realizou-se agrupamento dos dados sob o atributo 'aut_profissao'. Assim como no caso das vítimas, o atributo que descreve a profissão do autor conta com muitas categorias

distintas, fato que inviabiliza a apresentação de todas as profissões encontradas no *dataset*. Desse modo, optou-se por apresentar as 10 com maior número de registros, as quais são evidenciadas através do quadro 7.

QUADRO 7 – REGISTROS DE OCORRÊNCIAS POR PROFISSÃO DO AUTOR

Grau de instrução do autor	Registros
Desconhecido(a)	24.294
Desempregado	81
Serviços Gerais	21
Autônomo	20
Aposentado	8
Estudante	8
Comerciante	7
Vendedor	5
Pedreiro	3
Zelador	3

FONTE: A autora (2023).

Com o objetivo de entender quem são os autores de VCM presentes no *dataset* investigou-se o atributo ‘grau_de_relacionamento’, que descreve o envolvimento social entre vítimas e autores. Assim como nos atributos anteriores, a grande maioria dos registros contém valores nulos, contabilizando 22.816. Esses registros nulos foram somados às categorias ‘não informado’ e ‘vizinho’, gerando a categoria ‘Outros’, que conta com 23.170 registros. Ignorando os registros nulos, os graus de relacionamento que contam com maior número de ocorrências são ‘parceiro íntimo’ e ‘ex-parceiro íntimo’, com 713 e 400 registros, respectivamente, o que corresponde a 64,8% dos dados válidos. Embora a categoria ‘pai/padrasto’ represente o menor número de registros, conforme o quadro 8, ela representa 6,75% dos dados válidos.

QUADRO 8 – REGISTROS DE OCORRÊNCIAS POR GRAU DE RELACIONAMENTO COM A VÍTIMA

Grau de relacionamento	Registros
Outros	23.170
Parceiro íntimo	713
Ex-parceiro íntimo	400
Familiar	135
Pai/padrasto	116

FONTE: A autora (2023).

Mediante análise do atributo ‘meio_empregado’, explícita através do quadro 9, nota-se que a categoria com maior número de ocorrências é ‘Outros’, que agrupa valores como ‘fio de nylon’ e ‘corda’, além de registros nulos, e conta com 7.929 registros. A segunda categoria com maior número de ocorrências é ‘Homem’, com 5.656 registros, que agrupa todos os registros que originalmente continham essa palavra no atributo ‘meio_empregado’. De acordo com Costa (2022, é comum o valor “HOMEM/ADULTO” ser cadastrado em conjunto com outros meios, pois identifica que pelo menos um homem participou da ocorrência. A categoria com menor quantidade de ocorrências é a ‘automóvel’, com 198 registros, que agrupa os meios empregados "automovel", "motoneta", "meio de locomocao nao identificado", "motocicleta", "automovel de passeio", "onibus", "caminhao", "trator", "bicicleta", "automovel comercial".

QUADRO 9 – REGISTROS DE OCORRÊNCIAS POR MEIO EMPREGADO

Meio empregado	Registros
Outros	7.929
Homem	5.656
Sem instrumento	4.008
Ameaça	2.662
Agressão	1.632
Via digital	1.419
Armas em geral	1.030
Automóvel	198

FONTE: A autora (2023).

É possível destacar dentre os diferentes meios empregados a categoria ‘Via digital’, que é o agrupamento de valores como "telefone", "mensagem de texto", "celular", "whatsapp", "aplicativos de comunicação", "e-mail", "rede social" e "internet";

tal categoria representa um número expressivo de dados e pode ser associada ao crime de perseguição previsto na Lei 14.132/2021 (Brasil, 2021).

Em um próximo momento, buscou-se entender onde ocorreram os crimes de violência contra a mulher representados no *dataset*, conforme o quadro 10. Para tanto, explorou-se o atributo 'local_ocorrência', observando que, diferentemente dos atributos descritos acima, a maior parte dos dados possuem valor válido, sendo que 18.225 registros se referem a crimes cometidos em residências. Tal observação está em concordância com os dados relatados no 17º Anuário Brasileiro de Segurança Pública, que informou que 68,3% dos estupros e 69,3% dos feminicídios brasileiros ocorrem dentro de casa (FBSP, 2023).

QUADRO 10 – REGISTROS DE OCORRÊNCIAS POR LOCAL DE OCORRÊNCIA

Local ocorrência	Registros
Residência	18.225
Via pública	3.159
Outros	2.049
Estabelecimento	1.101

FONTE: A autora (2023).

Para finalizar a descrição estatística do conjunto de dados, analisou-se a contagem do atributo 'registros', que representa o tipo de violência praticada contra a mulher. Após o agrupamento dos 62 tipos de VCM distintos presentes no *dataset* em 9 categorias genéricas, conforme detalhado no quadro 11, observou-se que a categoria com maior expressividade é a 'Descumprir decisão judicial que defere medidas protetivas de urgência', com 11.781 registros, e a categoria com menor contagem é 'Feminicídio', com 316 ocorrências.

QUADRO 11 – REGISTROS DE OCORRÊNCIAS POR REGISTRO

Registro	Registros
Descumprir decisão judicial que defere medidas protetivas de urgência	11.781
Violência doméstica e familiar contra a mulher	3.070
Lesão corporal	2.878
Ameaça	2.621
Estupro	2.136
Outros	707
Importunação sexual	654
Tentativa de feminicídio	371
Feminicídio	316

FONTE: A autora (2023).

Os resultados encontrados na análise descritiva apresentada anteriormente podem ser comparados aos dados mais recentes sobre violência contra a mulher encontrados no 17º Anuário Brasileiro de Segurança Pública (FBSP, 2023). Embora o Anuário apresente estatísticas consolidadas para o feminicídio e o estupro, especificamente, é possível traçar um paralelo entre os resultados desta pesquisa, que considera todos os tipos de violência contra a mulher na análise descritiva, e os reportados pelo Anuário. Neste trabalho, a faixa etária com maior quantidade de ocorrência de VCM é a que compreende as idades entre 18 a 24 anos, em concordância com os dados apresentados pelo Anuário para a faixa etária com maior percentual de feminicídios em 2022. Quanto às vítimas de estupro, a faixa etária com maior incidência reportada pelo Anuário é a que compreende as idades entre 10 a 13 anos.

Quanto a cor da pele da vítima, nesta pesquisa observa-se maior frequência de vítimas brancas, em contraponto aos resultados de violência apresentados pelo Anuário, onde são evidenciadas taxas 30% maiores para vítimas negras de feminicídio e 16% maiores para vítimas negras de estupro em relação às vítimas brancas. Tal achado pode ser justificado pelo fato de que o estado do Paraná, que possui uma taxa maior de população branca, ser o com maior quantidade de registros na base de dados. Quanto ao horário com maior frequência de ocorrências de violência contra a mulher, nesta pesquisa observou-se maior quantidade de registros para o horário compreendido entre 18h e 00h, assim como o informado no 17º Anuário para o crime de estupro. Analisando o local de ocorrência dos crimes de VCM nesta base,

observou-se que a grande maioria ocorreu em residência, assim como os dados de feminicídio e estupro publicados pelo Anuário.

Após compreender os registros dos dados nos atributos mais relevantes presentes na base, foi possível definir o atributo 'registro' como atributo meta para a mineração de dados. Essa definição se justifica pela relevância que o atributo apresenta, visto que descreve as classificações possíveis para as ocorrências de violência contra a mulher, além de contar com alto número de dados válidos. Em suma, todo o processo de analisar descritivamente a base de ocorrências de VMC possibilitou conhecer com profundidade os dados, fator de extrema importância para definir os métodos e algoritmos de mineração de dados, cujos resultados são apresentados a seguir.

4.2 MINERAÇÃO DE DADOS

Nesta seção, são apresentados os experimentos realizados para definir a execução dos algoritmos Random Forest e Multilayer Perceptron no *software* Orange, bem como os resultados obtidos a partir do treinamento dos modelos. A seguir são apresentados os testes para definição das configurações aplicadas nos parâmetros de cada widget, para que então fosse possível definir aquelas a serem utilizadas no *workflow* final.

Após a leitura do *dataset* unificado pelo widget 'CSV Import File' e criação do atributo meta através do widget 'Create Class', o próximo passo foi configurar o widget 'Data Sampler', a fim de testar e definir os diferentes métodos de amostragem disponíveis no componente. Através dos resultados individuais das técnicas proporção fixa de dados, tamanho fixo de amostra, validação cruzada e *bootstrap*, optou-se por seguir para as configurações dos próximos widgets com os dois últimos métodos de amostragem, a validação cruzada e o *bootstrap*.

O método *cross validation* (validação cruzada) é uma técnica estatística eficiente para testes de desempenho de modelos de *machine learning* que divide aleatoriamente o conjunto de dados de treinamento em K subconjuntos, denominados *folds*. Desse modo, cada *fold* gerado é treinado e avaliado K vezes, gerando um modelo reduzido, que é submetido a avaliação e realização de previsões utilizando o modelo reduzido gerado a partir dos *folds* remanescentes. O *bootstrapping* (comumente chamado por *bootstrap*) é um método de que estima a distribuição de

amostragem ao coletar várias amostras com reposição de uma amostra aleatória única, dessa maneira obtendo reamostragem dos dados.

Definidos os métodos de amostragem a serem comparados, o próximo passo foi a testagem de três diferentes técnicas de seleção de atributos disponíveis no widget ‘Preprocess’ da ferramenta Orange. Com base nas técnicas de seleção de atributos estudadas por Costa (2022), foram escolhidas para teste as técnicas Info Gain, Gain Ratio e ReliefF.

Conforme informado nos encaminhamentos metodológicos, optou-se por manter as configurações padrão de parâmetros dos widgets dos algoritmos MLP e Random Forest. Desse modo, as três técnicas de seleção de atributos foram submetidas a testes utilizando cada uma das duas técnicas de amostragem para cada um dos algoritmos de classificação selecionados, resultando assim em quatro tabelas de resultados, conforme apresentados a seguir.

Ao executar o *workflow* para cada uma das combinações entre algoritmo, técnica de seleção de atributos e métodos de amostragem, no *widget* ‘Test and Score’, utilizando a opção de execução considerando os dados de teste, foram selecionadas para avaliação as métricas de tempo de processamento dos dados de treino e dos dados de teste, acurácia, precisão, *recall* e F1-Score. A métrica acurácia apresenta a proporção de classificação correta dos modelos. A métrica ‘precisão’ mede a proporção de instâncias positivas previstas corretamente em relação ao total de instâncias realmente positivas. A métrica ‘*recall*’ mede a capacidade do modelo de identificar todos os exemplos positivos no conjunto de dados. O F1-Score apresenta uma média ponderada entre as métricas precisão e *recall*.

A tabela 4 apresenta as métricas de avaliação do algoritmo Multilayer Perceptron para as três técnicas de seleção de atributos utilizando o método de amostragem *bootstrap*.

TABELA 4 – RESULTADOS DA EXECUÇÃO DO ALGORITMO MLP NO MÉTODO DE AMOSTRAGEM *BOOTSTRAP*

Técnica	Tempo de processamento dos dados de treino	Tempo de processamento dos dados de teste	Acurácia	Precisão	Recall	F1-Score
Info Gain	79,930	0,567	0,865	0,869	0,865	0,86
Gain Ratio	99,788	0,942	0,820	0,823	0,820	0,810

ReliefF	92,087	0,610	0,848	0,850	0,848	0,842
---------	--------	-------	-------	-------	-------	-------

FONTE: A autora (2023).

A tabela 5 apresenta as métricas de avaliação do algoritmo Multilayer Perceptron para as três técnicas de seleção de atributos utilizando o método de amostragem *cross validation*.

TABELA 5 – RESULTADOS DA EXECUÇÃO DO ALGORITMO MLP NO MÉTODO DE AMOSTRAGEM *CROSS VALIDATION*

Técnica	Tempo de processamento dos dados de treino	Tempo de processamento dos dados de teste	Acurácia	Precisão	Recall	F1-Score
Info Gain	80,035	0,671	0,871	0,875	0,871	0,867
Gain Ratio	92,510	0,874	0,823	0,828	0,823	0,814
ReliefF	79,750	0,632	0,861	0,864	0,861	0,857

FONTE: A autora (2023).

Em ambos os métodos de amostragem a técnica de seleção de atributos Gain Ratio é a que apresenta mais baixo desempenho na execução do modelo MLP, enquanto o Info Gain apresentou os melhores resultados em relação as outras duas técnicas. Embora as taxas da técnica Info Gain sejam muito semelhantes para ambos os métodos de amostragem, o *cross validation* apresentou acurácia, precisão, *recall* e F1-Score mais altos do que o método *bootstrap*, embora esse último tenha apresentado os menos tempos de processamento.

Em um próximo momento submeteu-se as mesmas técnicas de seleção de atributos para treinamento com o algoritmo Random Forest. A tabela 6 apresenta o resultado das métricas de avaliação do RF utilizando o método de amostragem *bootstrap*.

TABELA 6 – RESULTADOS DA EXECUÇÃO DO ALGORITMO RF NO MÉTODO DE AMOSTRAGEM *BOOTSTRAP*

Técnica	Tempo de processamento dos dados de treino	Tempo de processamento dos dados de teste	Acurácia	Precisão	Recall	F1-Score
Info Gain	1,045	0,510	0,917	0,922	0,917	0,914

Gain Ratio	1,498	0,797	0,834	0,838	0,834	0,827
ReliefF	2,093	0,485	0,929	0,933	0,929	0,927

FONTE: A autora (2023).

A tabela 7 apresenta as métricas de avaliação do algoritmo RF para as três técnicas de seleção de atributos utilizando o método de amostragem *cross validation*.

TABELA 7 – RESULTADOS DA EXECUÇÃO DO ALGORITMO RF NO MÉTODO DE AMOSTRAGEM *CROSS VALIDATION*

Técnica	Tempo de processamento dos dados de treino	Tempo de processamento dos dados de teste	Acurácia	Precisão	Recall	F1-Score
Info Gain	0,952	0,441	0,918	0,922	0,918	0,915
Gain Ratio	1,646	0,675	0,838	0,842	0,838	0,832
ReliefF	1,978	0,439	0,882	0,884	0,882	0,878

FONTE: A autora (2023).

Embora na execução do algoritmo RF a técnica Gain Ratio tenha apresentado taxas melhores do que no MLP, ela ainda é a que apresenta menor desempenho em relação as outras duas técnicas. Nesse modelo, é possível observar que para cada método de amostragem houve uma técnica de seleção de atributos que teve as melhores métricas de avaliação. Para o método *bootstrap*, a técnica ReliefF foi a que apresentou os melhores resultados. Já para o *cross validation*, a técnica Info Gain obteve as melhores taxas. De todo modo, visto que em ambos os modelos de *machine learning* a técnica de seleção de atributos Info Gain apresentou bons resultados, ela foi a escolhida para a execução do *workflow* final desta pesquisa. Em relação ao método de amostragem, na execução de ambos os algoritmos o *cross validation* obteve melhor desempenho.

Nesse contexto, a tabela 8 apresenta as métricas de avaliação para comparação entre os algoritmos MLP e RF utilizando o método de amostragem *cross validation* e a técnica de seleção de atributos Info Gain.

TABELA 8 – RESULTADOS DA EXECUÇÃO DOS ALGORITMOS RF E MLP

Técnica	Tempo de processamento dos dados de treino	Tempo de processamento dos dados de teste	Acurácia	Precisão	Recall	F1-Score
MLP	80,035	0,671	0,871	0,875	0,871	0,867
RF	0,952	0,441	0,918	0,922	0,918	0,915

FONTE: A autora (2023).

Ao comparar os algoritmos MLP e RF, pode-se perceber o quanto o modelo Random Forest se mostra superior para o *dataset* de violência contra a mulher construído para essa pesquisa. Em todas as métricas de avaliação o resultado do RF supera as métricas do MLP. De todo modo, é possível perceber que ambos os algoritmos apresentaram altas taxas de acurácia.

Tendo definido o método de amostragem *cross validation* e a técnica de seleção de atributos como os parâmetros de configuração do *workflow*, foi possível gerar as matrizes de confusão para ambos os algoritmos empregados nesta pesquisa. A figura 21 apresenta a matriz de confusão do algoritmo MLP.

FIGURA 21 – MATRIZ DE CONFUSÃO DO ALGORITMO MLP

		Predicted									
		estupro	descu...	ameaca	import...	violenc...	lesao_c...	tentati...	outros feminic...	Σ	
Actual	estupro	92.0 %	0.0 %	0.0 %	8.2 %	0.0 %	0.0 %	0.0 %	6.8 %	2149	
	descu...	0.0 %	94.9 %	3.0 %	0.0 %	5.4 %	0.4 %	0.8 %	0.4 %	11716	
	ameaca	0.0 %	0.0 %	67.1 %	0.0 %	0.0 %	12.7 %	10.1 %	2.8 %	2706	
	import...	4.8 %	0.0 %	0.0 %	88.8 %	0.0 %	0.0 %	0.0 %	5.1 %	657	
	violenc...	0.0 %	5.0 %	0.0 %	0.0 %	94.6 %	0.0 %	0.0 %	0.0 %	3018	
	lesao_c...	0.0 %	0.1 %	17.8 %	0.0 %	0.0 %	80.9 %	3.8 %	2.3 %	2863	
	tentati...	0.0 %	0.0 %	3.5 %	0.0 %	0.0 %	1.8 %	76.9 %	1.1 %	365	
	outros	3.2 %	0.0 %	4.0 %	3.1 %	0.0 %	2.2 %	5.9 %	80.5 %	731	
	feminic...	0.0 %	0.0 %	4.7 %	0.0 %	0.0 %	1.9 %	2.5 %	1.1 %	329	
Σ		2244	12073	3424	588	2555	2727	238	532	153	24534

FONTE: A autora (2023).

Na figura 22, é possível visualizar a matriz de confusão do algoritmo Random Forest.

FIGURA 22 – MATRIZ DE CONFUSÃO DO ALGORITMO RF

		Predicted									Σ
		estupro	descu...	ameaca	import...	violenc...	lesao_c...	tentati...	outros	feminic...	
Actual	estupro	88.4 %	0.0 %	0.0 %	2.3 %	0.0 %	0.1 %	0.0 %	0.7 %	0.0 %	2149
	descu...	0.0 %	97.0 %	2.6 %	0.0 %	3.0 %	0.6 %	0.0 %	0.0 %	2.0 %	11716
	ameaca	0.0 %	0.0 %	79.5 %	0.0 %	0.0 %	6.2 %	1.4 %	0.2 %	0.0 %	2706
	import...	7.3 %	0.0 %	0.0 %	92.6 %	0.0 %	0.0 %	0.0 %	0.4 %	0.0 %	657
	violenc...	0.0 %	3.0 %	0.0 %	0.0 %	97.0 %	0.0 %	0.0 %	0.0 %	0.0 %	3018
	lesao_c...	0.0 %	0.0 %	7.8 %	0.0 %	0.0 %	87.5 %	1.8 %	0.4 %	0.0 %	2863
	tentati...	0.0 %	0.0 %	3.3 %	0.0 %	0.0 %	1.6 %	93.6 %	0.0 %	4.1 %	365
	outros	4.2 %	0.0 %	2.9 %	5.0 %	0.0 %	1.9 %	2.7 %	97.6 %	0.0 %	731
	feminic...	0.0 %	0.0 %	3.9 %	0.0 %	0.0 %	2.1 %	0.5 %	0.7 %	93.9 %	329
Σ		2409	11894	3169	516	2738	2983	220	457	148	24534

FONTE: A autora (2023).

A matriz de confusão deve ser interpretada da seguinte maneira: as linhas representam as categorias reais do atributo meta presente na base, portanto a última coluna visível apresenta a contagem total real de cada categoria. Por outro lado, as colunas representam a porcentagem de predição do modelo e, nesse sentido, a última linha visível apresenta a contagem total de instâncias preditas para cada categoria.

Apresentados os resultados obtidos através do processamento dos modelos de *machine learning* Multilayer Perceptron e Random Forest, a próxima seção apresenta uma análise crítica sobre a análise de dados criminais de violência contra a mulher.

4.3 ANÁLISE CRÍTICA SOBRE A ANÁLISE DE DADOS CRIMINAIS DE VIOLÊNCIA CONTRA A MULHER

Em concordância com os pensamentos de Okabe e Fonseca (2009), destaca-se a importância da visibilidade do fenômeno da violência contra a mulher para a superação das relações de desigualdade, que pode ser conseguida por meio da informação à população em vistas a romper a indiferença e a cumplicidade que levam à naturalização e à permanência de práticas discriminatórias na sociedade. Nesse sentido, ampliar o conhecimento do fenômeno através da análise de dados promove o reconhecimento da desigualdade e da forma como ela evolui, fundamentando as intervenções para a mudança.

Para Campos e Almeida (2017), a produção de informação acerca da violência de gênero pode alterar relações de poder e gerar conhecimentos úteis para estimular pesquisas e estratégias de políticas públicas que propiciem mudanças no âmbito social, político e cultural no que se refere aos direitos das mulheres. Dessa maneira, conduzindo a uma perspectiva direcionada à igualdade de gênero e à justiça social, potencializando a resistência frente as mais distintas desigualdades. Nesse contexto, os autores reforçam que o uso de sistemas de informação possibilita o armazenamento e processamento de dados e a disponibilização de informações, tornando-as acessíveis e úteis não apenas à administração pública e seus processos organizacionais, mas também adequadas e voltadas à sociedade de uma forma geral.

Choo (2003), abordando a importância da gestão da informação sob a perspectiva da Gestão Pública, discorre sobre a crescente complexidade do cenário que envolve constantes demandas que se desdobram dos anseios da sociedade. Para o autor, a contribuição da gestão da informação é ser inserida no desenvolvimento de instrumentos de gestão das rotinas organizacionais somadas ao tratamento de dados brutos para o uso estratégico da informação pelo poder público em várias áreas sociais de atuação, colaborando para dar sentido às mudanças no ambiente externo à organização, no caso, a sociedade.

Para Monteiro e Caballero (2021), o sistema de informação e estatísticas da área de Segurança Pública no Brasil possui problemas e limitações, sobretudo quando comparado a sistemas de informações de outras áreas centrais de provisão de serviços públicos, como saúde e educação. As autoras chamam especial atenção para o fato de que grande parte das estatísticas são advindas de sistemas de informações estaduais, fato que exige cautela na comparação dos números. Nesse sentido, afirmam que é preciso entender o que cada número está representado para a comparação de maneira correta, visto que é possível cair em armadilhas ao analisar estados com características sociais diferentes.

Monteiro e Caballero (2021) acrescentam que um problema de relevância em sistemas de informação limitados é a necessidade de despender grandes esforços para obter o mínimo de informação sobre o tema, influenciando na falta de eficiência no uso dos recursos públicos e na redução da violência no país. De acordo com as autoras, o primeiro passo crucial para a identificação de prioridades e a qualificação do problema no desenvolvimento de trabalhos e pesquisas, que podem contribuir significativamente para a avaliação do cenário, consiste em possuir dados confiáveis

e comparáveis. Esses dados são essenciais na construção de evidências que, por sua vez, têm o potencial de influenciar políticas voltadas para a promoção da Segurança Pública no Brasil.

Um achado de relevância nessa pesquisa é a ausência significativa de dados válidos sobre as características demográficas dos autores de ocorrências de violência contra a mulher. Essa falta de dados pode ter várias implicações importantes na análise e interpretação dos resultados da pesquisa, como viés na caracterização do perfil dos perpetradores desse tipo de crime, pois ter poucos dados sobre idade, escolaridade, ocupação e outros aspectos demográficos pode limitar a compreensão dos fatores associados à violência contra a mulher, comprometendo a capacidade de identificar correlação entre as variáveis e descoberta de padrões específicos.

No desenvolvimento dessa pesquisa, tomou-se conhecimento da Lei 14.232/2021, que institui a Política Nacional de Dados e Informações relacionadas à Violência contra as Mulheres (PNAINFO). Seu propósito principal é a coleta, organização e disponibilização de dados sobre diversas formas de violência contra a mulher, buscando integrar as bases de órgãos do Executivo, Legislativo e Judiciário que atendem mulheres em situação de violência. Além de fomentar a transparência na gestão dessas informações, a política visa incentivar a participação social por meio da contribuição de dados para a avaliação das políticas públicas de combate à violência de gênero. Suas diretrizes destacam a integração de bases de dados, transparência na produção de informações e promoção da participação social (Brasil, 2021).

As metas da PNAINFO englobam subsidiar a formulação e implementação de políticas, gerar informações fidedignas sobre violência contra mulheres, manter dados acessíveis eletronicamente, integrar e respaldar políticas específicas, cumprir legislações vigentes, padronizar e disponibilizar indicadores, bem como atender a acordos internacionais relacionados à produção de dados sobre violência contra mulheres. A legislação estabelece a criação de um Registro Unificado de Dados e Informações sobre Violência contra as Mulheres em meio eletrônico, abrangendo registros administrativos, serviços especializados e políticas públicas.

A elaboração dessa lei é um importante passo na superação das limitações e desafios atrelados à análise de dados sobre a violência contra a mulher. Entretanto, a fiscalização para que seu cumprimento seja atingido deve ser constante, aliada a ações como a capacitação, sensibilização e aperfeiçoamento do pessoal responsável

pelo preenchimento das ocorrências e possibilidade de colaboração entre profissionais de variadas áreas do conhecimento a fim de identificar possibilidades de melhoria em políticas públicas.

5 CONSIDERAÇÕES FINAIS

Diante do exposto nas seções anteriores, fica evidente a necessidade de decisões baseadas em dados para uma compreensão abrangente sobre a violência de gênero no país e para a implementação de políticas eficazes de prevenção e combate a esse grave problema social. Entretanto, a análise de dados sobre o fenômeno enfrenta várias limitações e desafios, como a falta de padronização na coleta e categorização dos dados, além de questões relacionadas à privacidade e à ética, que influenciam na disponibilidade dos dados. Para superá-los é essencial adotar abordagens multidisciplinares e colaborativas entre instituições governamentais, organizações não governamentais, pesquisadores e a sociedade civil.

Visando analisar a violência de gênero à luz dos conhecimentos da Gestão da Informação, esta pesquisa se valeu de dois campos caros à disciplina, a ciência de dados e a ciência social. O problema de pesquisa proposto na introdução norteou o percurso deste trabalho e, através de seu desenvolvimento, foi possível vivenciar o panorama que compõe as limitações e desafios da análise de dados criminais sobre violência contra a mulher. Durante a aplicação de cada etapa do processo KDD que orientou os caminhos metodológicos, pode-se lidar com diferentes dilemas e impasses ao analisar dados deste fenômeno social.

Vale lembrar que a disponibilidade de dados provenientes das ocorrências criminais de violência contra a mulher de dois estados brasileiros nesta pesquisa é um fator de grande relevância, visto a dificuldade de acesso de dados dessa natureza e complexidade. Entretanto, observou-se que essa é a primeira limitação encontrada pelos pesquisadores em geral; nesse sentido, pesquisas com dados dos sistemas de saúde compõe a grande maioria da produção científica sobre a VCM.

Outra dificuldade encontrada foi a falta de padronização de coleta e disponibilização dos dados. Embora existam entre as duas bases um número considerável de atributos correlatos, ainda é possível encontrar diferentes padrões de preenchimento e valores disponíveis nos registros. Nesse sentido, têm-se uma limitação de comparação entre dados de regiões tão diferentes do Brasil tornando as possibilidades de interpretação de diferenças de padrão tão significativas quanto as diferenças geográficas. De todo modo, o fato de conseguir **construir uma base**

unificada que reúne metadados sobre a VCM nos estados Pará e Paraná satisfaz o primeiro objetivo específico proposto na introdução desta pesquisa.

Tendo a base construída, os próximos desafios foram as tomadas de decisão quando à limpeza de dados, agrupamento de categorias, tratamento de dados faltantes, agregação e deleção de colunas. Nesse sentido, realizar os tratamentos da etapa de pré-processamento possibilitou conhecer os dados com mais profundidade e, então, definir as análises que foram desenvolvidas, tais quais: estatística descritiva, mineração de dados e visualizações de registros dos dados. Para tanto, foram utilizados *softwares* e linguagem de programação que permitiram o tratamento, a análise descritiva e a mineração dos dados, que se deu através do treinamento dos algoritmos Random Forest e Multilayer Perceptron. Desse modo, ao **estabelecer a metodologia de análise, definir a ferramenta de mineração de dados, utilizar e comparar os métodos** permitiu o atingimento do segundo objetivo específico proposto.

A expressiva ausência de dados que descrevem os autores de VCM na base construída pode ser considerada uma limitação relevante, pois influencia na descoberta de padrões e na caracterização dos indivíduos que perpetram esse tipo de crime. De todo modo, a utilização da ferramenta Orange para a mineração dos dados possibilitou testar e comparar diferentes algoritmos de classificação, métodos de amostragem e técnicas de seleção de atributos, para **então avaliar os resultados obtidos por meio de critérios estatísticos** como a acurácia, precisão e recall, e assim, atingir o terceiro objetivo específico proposto.

O Anuário Brasileiro de Segurança Pública representa uma rica fonte de informação para conhecimento do panorama da violência de gênero no país, portanto, pode-se destacar sua importância para comparação dos resultados obtidos na análise descritiva dos dados deste trabalho. Ainda, pode-se citar as contribuições referenciais de autores como Campos e Almeida (2017) e Monteiro e Caballero (2021) para a elaboração da análise crítica sobre a análise de dados sobre violência contra a mulher. Nesse sentido, atingiu-se o último objetivo específico proposto na definição do problema desta pesquisa, ao **alinhar os resultados obtidos com o levantamento de literatura e produzir *insights* de aplicação**.

A pesquisa destaca a importância de abordar não apenas a coleta e análise de dados, mas também os desafios sociais, culturais e éticos associados à violência contra a mulher. Superar essas limitações exige esforços colaborativos, inovação e

um compromisso contínuo com a promoção da igualdade e segurança para todas as mulheres. Os resultados encontrados, seja os advindos da análise descritiva ou da análise crítica proposta, são a grande contribuição deste trabalho. A informação de que aproximadamente 48% dos tipos de violência cometido contra a mulher nesta base de dados são referentes ao descumprimento de medida protetiva de urgência é um fato que pode contribuir para balizar ações públicas de enfrentamento à essa questão, seja ações preventivas ou punitivas.

Nesse sentido, ao elencar as dificuldades e desafios encontrados na análise de dados criminais sobre a violência contra a mulher, responde-se à pergunta de pesquisa proposta. De maneira complementar, ao atingir os objetivos específicos, além de discutir os resultados obtidos, se atinge o objetivo geral proposto de **realizar uma análise crítica sobre a análise de dados criminais de violência contra a mulher em estados brasileiros**.

Cabe, por fim, enfatizar que a pesquisa não extenua os conhecimentos desta problemática, apenas aborda algumas particularidades da realidade da análise de dados criminais de violência contra a mulher. É possível elencar como limitação desta pesquisa a experiência para a construção do *workflow* na ferramenta Orange, visto que demanda conhecimentos tanto sobre as técnicas de mineração quanto conhecimento aprofundado sobre os dados e sobre a problemática que eles representam.

Dessa forma, novas pesquisas que se debrucem sobre o tema tornam-se imprescindíveis. Recomenda-se que, para trabalhos futuros, é possível explorar melhor os diversos componentes, e combinações possíveis entre eles, do software Orange, além de abordagem madura sobre a manipulação de dados desse fenômeno, considerando questões éticas e metodológicas sobre a imputação de registros nulos, o estudo de técnicas como *clustering* para agrupamento de categorias com muitos valores distintos, além de compreensão aprofundada sobre as métricas de avaliação dos modelos.

REFERÊNCIAS

ANJOS JUNIOR, O. R. dos; PORCINO, M. M. Violência contra mulheres na Paraíba: um enfoque para os casos de homicídios no período de 2011 a 2017. **Revista Brasileira de Segurança Pública**, São Paulo, v. 15, n. 1, p. 74-91, 2021. Disponível em: <https://revista.forumseguranca.org.br/index.php/rbsp/article/view/1110>. Acesso em: 31 ago. 2023.

ARAÚJO, E. A. de. Informação, sociedade e cidadania: gestão da informação no contexto de organizações não-governamentais (ONGs) brasileiras. **Ciência da Informação**, Brasília, v. 29, n. 2, p. 155-167, 1999. Disponível em: <https://www.scielo.br/j/ci/a/hBLTc6jXX6s3C4QDtnjGXG/abstract/?lang=pt>. Acesso em: 15 maio 2023.

BARBOSA, R. R. Gestão da informação e do conhecimento: origens, polêmicas e perspectivas. **Informação & Informação**, v. 13, n. esp., p. 1-25, 2008. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/1843>. Acesso em: 15 maio 2023.

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: 29 ago. 2023.

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://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13104.htm. Acesso em: 17 abr. 2023.

BRASIL. Lei nº 13.772, de 19 de dezembro de 2018. Altera a Lei nº 11.340 e o Decreto-Lei nº 2.848, de 7 de dezembro de 1940 - Código Penal, para reconhecer que a violação da intimidade da mulher configura violência doméstica e familiar e para criminalizar o registro não autorizado de conteúdo com cena de nudez ou ato sexual ou libidinoso de caráter íntimo e privado. **Diário Oficial da União**, Brasília, DF, n. 244, 20 dez. 2018. Seção 1, p. 1. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13772.htm. Acesso em: 17 abr. 2023.

BRASIL. Lei nº 14.132, de 31 de março de 2021. Acrescenta o art. 147-A ao Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), para prever o

crime de perseguição; e revoga o art. 65 do Decreto-Lei nº 3.688, de 3 de outubro de 1941 (Lei das Contravenções Penais). **Diário Oficial da União**, Brasília, DF, n. 244, 01 abr. 2021. Seção 1, p. 1. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14132.htm. Acesso em: 18 nov. 2023.

BRASIL. Lei nº 14.232, de 28 de outubro de 2021. Institui a Política Nacional de Dados e Informações relacionadas à Violência contra as Mulheres (PNAINFO). **Diário Oficial da União**, Brasília, DF, n. 244, 29 out. 2021. Seção 1, p. 1. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14232.htm. Acesso em: 25 nov. 2023.

BREIMAN, L. **Random forests**, 2001. Disponível em: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>. Acesso em: 20 nov. 2023.

CAICEDO-ROA, M.; BANDEIRA, L. M.; CORDEIRO, R. C. Femicídio e feminicídio: discutindo e ampliando os conceitos. **Revista Estudos Feministas**, Florianópolis, v. 30, n. 3, e83829, 2022. Disponível em: <https://periodicos.ufsc.br/index.php/ref/article/view/83829>. Acesso em: 24 abr. 2023.

CAICEDO-ROA, M.; CORDEIRO, R. C. Análise de casos de feminicídio em Campinas, SP, Brasil, entre 2018 e 2019 por meio do modelo ecológico da violência. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 28, n. 1, p. 23-36, 2023. Disponível em: <https://www.scielo.br/j/csc/a/mRzcdz4kyz4rsFtsf8gQ3RB/abstract/?lang=pt>. Acesso em: 28 mar. 2023.

CAMPOS, M. L.; ALMEIDA, G. H. M. D. Violência contra a mulher: uma relação entre dimensões subjetivas e a produção de informação. **Revista Digital de Biblioteconomia & Ciência da Informação**, Campinas, v. 15, n. 2, p. 349-367, 2017. Disponível em: <https://brapci.inf.br/index.php/res/v/40104>. Acesso em: 03 set. 2023.

CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva, 2016.

CHOO, C. W. A organização do conhecimento II: como equilibrar tensões e administrar a informação. In: CHOO, C. W. **A organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões**. São Paulo: Editora Senac São Paulo, 2003. p. 381-421.

CONEGLIAN, C. S. *et al.* Orange Data Mining: ferramenta para inserção de Inteligência Artificial na pesquisa científica. In: SHINTAKU, M.; MACÊDO, D. J.; MARIN, L. H. G. (orgs.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas**. Brasília, IBICT, p. 245-275, 2023.

COSTA, C. A. da S. **Proposta de fluxo de análise de dados preditiva em violência doméstica contra a mulher**. 2022. 121 f. Dissertação (Mestrado em Gestão da Informação) – Programa de Pós-Graduação em Gestão da Informação,

Universidade Federal do Paraná, Curitiba, 2022. Disponível em: <https://acervodigital.ufpr.br/handle/1884/80754>. Acesso em: 04 fev. 2023.

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

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 3, n. 17, p.37-54,1996. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>. Acesso em: 23 maio 2023.

FERRAZ, M. I. R. **Perfil da violência contra a mulher em Guarapuava-PR**. 2008. 86f. Dissertação (Mestrado em Enfermagem) – Programa de Pós-Graduação em Enfermagem, Universidade Federal do Paraná, Curitiba, 2008. Disponível em: <https://acervodigital.ufpr.br/handle/1884/18472>. Acesso em: 03 set. 2023.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA (FBSP). **17º Anuário Brasileiro de Segurança Pública**. São Paulo: Fórum Brasileiro de Segurança Pública, 2023. Disponível em: <https://forumseguranca.org.br/wp-content/uploads/2023/07/anuario-2023.pdf>. Acesso em: 29 jul. 2023.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA (FBSP). **Violência contra meninas e mulheres no 1º semestre de 2023**. São Paulo: FBSP, 2023. <https://forumseguranca.org.br/wp-content/uploads/2023/11/violencia-contra-meninas-mulheres-2023-1sem.pdf>. Acesso em: 18 nov. 2023.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA (FBSP). **Visível e invisível: a vitimização de mulheres no brasil**. São Paulo: FBSP, 2022. Disponível em: <https://forumseguranca.org.br/wp-content/uploads/2023/03/visiveleinvisivel-2023-relatorio.pdf>. Acesso em: 7 set. 2023.

GALVÃO, N. D.; MARIN, H. de F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, São Paulo, v. 22, n. 5, p. 686-690, 2009. Disponível em: <https://www.scielo.br/j/ape/a/Lzj9vW6Fp4QVdXyNKhmtfvv/abstract/?lang=pt>. Acesso em: 12 jun. 2023.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining**: conceitos, técnicas, algoritmos, orientações e aplicações. 2. ed. Rio de Janeiro: Elsevier, 2015.

LAGARDE, M. Del femicidio al feminicidio. **Desde el Jardín de Freud**, v. 6, p. 216-225, 2006. Disponível em: <https://revistas.unal.edu.co/index.php/jardin/article/view/8343>. Acesso em: 04 maio 2023.

LAMBERT, K. A. **Fundamentos de Python**: primeiros programas. São Paulo: Cengage Learning, 2022.

LUCENA, M. B. N. de. **Morte violenta de mulheres no brasil e novas vulnerabilidades**: da violência do patriarcado privado à violência do patriarcado público. 2020. 243 f. Tese (Doutorado em Ciências Criminais) – Programa de Pós-Graduação em Ciências Criminais, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2020. Disponível em: https://repositorio.pucrs.br/dspace/bitstream/10923/17171/6/TES_MARIANA_BARRETO_NOBREGA_DE_LUCENA_COMPLETO.pdf. Acesso em: 29 ago. 2023.

MARTINELLI, A. Violência contra a mulher: uma abordagem histórica. **Teoria Jurídica Contemporânea**, v. 5, n. 2, p. 12-44, 2020. Disponível em: <https://revistas.ufrj.br/index.php/rjur/index>. Acesso em: 08 set. 2023.

MENEGHEL, S. N.; HIRAKATA, V. N. Femicídios: homicídios femininos no Brasil. **Revista de Saúde Pública**, São Paulo, v. 45, n. 3, p. 564-574, 2011. Disponível em: <https://www.scielo.br/j/rsp/a/C6XjntCBHFNFjXZJ96tGMBN/abstract/?lang=pt>. Acesso em: 11 maio 2023.

MIRANDA, C. M.; CARVALHO, C. A. de. Narrativas do feminicídio na Amazônia. **Revista Estudos Feministas**, Florianópolis, v. 30, n. 2, e76976, 2022. Disponível em: <https://periodicos.ufsc.br/index.php/ref/article/view/76976>. Acesso em: 24 abr. 2023.

MONTEIRO, J.; CABALLERO, B. Crimes e violência. In: SHIKIDA, C. D.; MONASTERIO, L.; NERY, P. F. **Guia brasileiro de análise de dados**: armadilhas & soluções. Brasília: Enap, p. 127-169, 2021.

NETTO, A.; MACIEL, F. **Python para data science e machine learning descomplicado**. Rio de Janeiro: Alta Books, 2021.

NONATO, R. D. S.; AGANETTE, E. C. Gestão da informação: rumo a uma proposta de definição atual e consensual para o termo. **Perspectivas em Ciência da Informação**, v. 27, n. 1, 2022. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/203838>. Acesso em: 15 maio 2023.

OKABE, I.; FONSECA, R. M. G. S. da. Violência contra a mulher: contribuições e limitações do sistema de informação. **Revista da Escola de Enfermagem da USP**, v. 43, n. 2, p. 453-458, 2009. Disponível em: <https://www.revistas.usp.br/reeusp/article/view/40378>. Acesso em: 14 set. 2023.

PEREIRA, I. S. da S. D. *et al.* Feminicídio: um problema de saúde pública. **Revista Interface**, Natal, v. 19, n. 2, p. 114-131, 2022. Disponível em: <https://ojs.ccsa.ufrn.br/index.php/interface/article/view/1286>. Acesso em: 03 abr. 2023.

RADFORD, J.; RUSSELL, D. (Eds.). **Femicide**: the politics of woman killing. Nova York: Twayne Publishers, 1992.

ROICHMAN, C. B. C. Faca, peixeira, canivete: uma análise da lei do feminicídio no Brasil. **Revista Katálisis**, Florianópolis, v. 23, n. 2, p. 357-265, 2020. Disponível em:

<https://www.scielo.br/j/rk/a/7zQRkyKBpyYKHP6JXbKXrPr/?lang=pt>. Acesso em: 03 abr. 2023.

ROMEIRO, N. L.; BEZERRA, A. C. A naturalização da violência contra a mulher e a trajetória da criminalização da violência sexual no Brasil. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 13, n. 1, 2020. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/159390>. Acesso em: 18 abr. 2023.

ROMIO, J. A. F. **Feminicídios no Brasil, uma proposta de análise com dados do setor de saúde**. 2017. 215 f. Tese (Doutorado em Demografia) – Instituto de Filosofia e Ciências Humanas, Universidade Estadual de Campinas, Campinas, 2017. Disponível em: <http://repositorio.unicamp.br/Acervo/Detalhe/988584>. Acesso em: 03 abr. 2023.

RUSSELL, D. **The origin and importance of the term femicide**. Nova York, 2011. Disponível em: http://www.dianarussell.com/origin_of_femicide.html. Acesso em: 03 maio 2023.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. del P. B. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SANTIAGO, I. M. F. L. Violência de gênero entre usuárias do serviço de atenção básica do SUS na Paraíba. **Revista Katálisis**, Florianópolis, v. 24, n. 2, p. 386-396, 2021. Disponível em: <https://www.scielo.br/j/rk/a/WqshfBQtDdxYM7jPyqBPMpv/?lang=pt>. Acesso em: 14 set. 2023.

SANTOS, C. M.; IZUMINO, W. P. Violência contra as mulheres e violência de gênero: notas sobre estudos feministas no Brasil. **Estudios Interdisciplinarios de América Latina y el Caribe**, v. 16, n. 1, 2005. Disponível em: <https://eial.tau.ac.il/index.php/eial/article/view/482>. Acesso em: 17 set. 2023.

SCHERER, Z. A. P. *et al.* Feminicídio: estudo com dados do setor saúde de um município paulista. **Research, Society and Development**, v. 11, n. 9, e48611932013, 2022. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/32013>. Acesso em: 03 abr. 2023.

SHARDA, R.; DELEN, D.; TURBAN, E. **Business intelligence e análise de dados para gestão do negócio**. Porto Alegre: Bookman, 2019. *E-book*.

SILVA, L. E. L. da; OLIVEIRA, M, L. C. de. Violência contra a mulher: revisão sistemática da produção científica nacional no período de 2009 a 2013. **Ciência & Saúde Coletiva**, v. 20, n. 11, p. 3523-3532, 2015. Disponível em: <https://cienciaesaudecoletiva.com.br/artigos/violencia-contra-a-mulher-revisao-sistemica-da-producao-cientifica-nacional-no-periodo-de-2009-a-2013/15074?id=15074>. Acesso em: 24 set. 2023.

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

STEINER, M. T. A. *et al.* Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Gestão e Produção**, São Carlos, v.13, n.2, p.325-337, 2006. Disponível em: <https://www.scielo.br/j/gp/a/gC9RkgLD8B8FffPsNhBYWkB/abstract/?lang=pt>. Acesso em: 12 jun. 2023.

TARAPANOFF, K. Informação, conhecimento e inteligência em corporações: relações e complementaridade. *In*: TARAPANOFF, K (org.). **Inteligência, informação e conhecimento**. Brasília: IBICT, UNESCO, p. 19-35, 2006.

TOLOSA, T. da S. R.; CHAGAS, C. A. N.; LIMA, C. do S. L. Caracterização do feminicídio na cidade de Belém (PA). **Revista Brasileira de Segurança Pública**, São Paulo, v. 14, n. 2, p. 162-177, 2020. Disponível em: <https://revista.forumseguranca.org.br/index.php/rbsp/article/view/804>. Acesso em: 18 abr. 2023.

TSUNODA, D. F.; CONSTÂNCIO, A. S. Python como suporte às pesquisas sociais. *In*: SHINTAKU, M.; MACÊDO, D. J.; MARIN, L. H. G. (orgs.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas**. Brasília, IBICT, p. 61-89, 2023.

UNODC; UN WOMEN. **Gender-related killings of women and girls (femicide/feminicide)**: global estimates of gender-related killings of women and girls in the private sphere in 2021. 2022. Disponível em: https://www.unodc.org/documents/data-and-analysis/briefs/Femicide_brief_Nov2022.pdf. Acesso em: 15 maio 2023.

WASELFISZ, J. J. **Mapa da Violência 2015**: homicídio de mulheres no Brasil. Brasília, 2015. Disponível em: http://www.onumulheres.org.br/wp-content/uploads/2016/04/MapaViolencia_2015_mulheres.pdf. Acesso em: 25 maio 2023.

WORLD HEALTH ORGANIZATION (WHO). **Violence against women prevalence estimates, 2018**: global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women Geneva: World Health Organization, 2021. Disponível em: <https://www.who.int/publications/i/item/9789240022256>. Acesso em: 08 set. 2023.