Research, Society and Development, v. 12, n. 1, e20112139561, 2023
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v12i1.39561

Análise de dados de violência doméstica contra a mulher

Data analysis on domestic violence against the woman

Analisís de datos de violencia doméstica contra las mujeres

Recebido: 20/12/2022 | Revisado: 03/01/2023 | Aceitado: 05/01/2023 | Publicado: 08/01/2023

Christian Aristóteles da Silva Costa
ORCID: https://orcid.org/0000-0002-5309-9716
Universidade Federal do Paraná, Brasil
E-mail: christianascosta@gmail.com
Denise Fukumi Tsunoda
ORCID: https://orcid.org/0000-0002-5663-4534
Universidade Federal do Paraná, Brasil
E-mail: dtsunoda@ufpr.br

Resumo
O feminicídio é o pior desfecho de uma ocorrência nos casos de violência doméstica, pois a mulher falece depois sofrer violência uma ou mais vezes. As bases de dados sobre violência doméstica e feminicídio no estado do Paraná são compostas de muitos atributos e valorá-los para sua posterior análise é um problema que, sem o auxílio de um fluxo ou método, pode ser uma tarefa demorada e ineficaz. Objetivou-se nesse estudo construir fluxos de análise preditiva que apoiassem pesquisas com dados de segurança pública. Trata-se de uma pesquisa com abordagem quantitativa, método indutivo, de nível exploratório. A análise das bases de dados de violência contra à mulher dos anos de 2018, 2019 e 2020 foi realizada por meio da estatística descritiva combinada com o modelo de Fayyad para a descoberta de conhecimento por meio da mineração de dados que empregou quatro técnicas de seleção de atributos de abordagem filtro e algoritmos de indução de regras PRISM e CN2. A idade média é de 37 anos e a ocupação mais frequente está ligada ao serviço doméstico para as vítimas de ambas as bases de dados, 63% das mulheres mortas por feminicídio tem histórico de violência doméstica sendo mais provável que a vítima e o autor coabitem e 19% das vítimas registraram mais de uma ocorrência. As regras geradas pelo algoritmo CN2 com as técnicas de seleção de atributos CFS e Info Gain foram validadas por especialistas em análise criminal.
Palavras-chave: Análise de dados; Violência doméstica; Análise preditiva; Descoberta de conhecimento em bases de dados.

Abstract
Femicide is the worst outcome of a police occurrence in cases of domestic violence, as the woman dies after suffering violence one or more times. The databases on domestic violence and femicide in the state of Paraná are made up of many attributes and valuing them for further analysis is a problem that, without the aid of a flow or method, can be a time-consuming and ineffective task. The objective of this study was to build predictive analysis flows that support research with public safety data. This is a research with a quantitative approach, inductive method, at an exploratory level. The analysis of the violence against women databases for the years 2018, 2019 and 2020 was carried out using descriptive statistics combined with the Fayyad model for knowledge discovery through data mining that employed four attribute selection techniques with filter approach and rules induction algorithms PRISM and CN2. The average age is 37 years and the most frequent occupation is linked to domestic service for the victims of both databases, 63% of the women killed by femicide have a history of domestic violence, being more likely that the victim and the perpetrator cohabit and 19% of the victims registered more than one occurrences. The rules generated by the CN2 algorithm with the CFS and Info Gain attribute selection techniques were validated by specialists in criminal analysis.
Keywords: Data analysis; Domestic violence; Predictive analytics; Discovery of knowledge in databases.

Resumen
El feminicidio es el peor desenlace de una ocurrencia en los casos de violencia doméstica, ya que la mujer muere después de sufrir violencia una o más veces. Las bases de datos sobre violencia doméstica y feminicidio en el estado de Paraná están compuestas por muchos atributos y valorarlos para su posterior análisis es un problema que, sin la ayuda de un flujo o método, puede ser una tarea lenta e ineficaz. El objetivo de este estudio fue construir flujos de análisis predictivo que apoyen la investigación con datos de seguridad pública. Se trata de una investigación con enfoque cuantitativo, método inductivo, a nivel exploratorio. El análisis de las bases de datos de violencia contra las mujeres de los años 2018, 2019 y 2020 se realizó utilizando estadística descriptiva combinada con el modelo Fayyad para el descubrimiento de conocimiento a través de la minería de datos que empleó cuatro técnicas de selección de atributos de enfoque filtro y algoritmos de inducción para reglas PRISM y CN2. La edad promedio es 37 años y la ocupación más frecuente está ligada al servicio doméstico para las víctimas de ambas las bases de datos, el 63% de las

mujeres muertas por femicidio tienen antecedentes de violencia doméstica, siendo más probable que la víctima y el victimario cohabiten y 19 % de las víctimas registraron más de una ocurrencia. Las reglas generadas por el algoritmo CN2 con las técnicas de selección de atributos CFS e Info Gain fueron validadas por especialistas en análisis criminal.
Palabras clave: Análisis de datos; La violencia doméstica; Análisis predictivo; Descubrimiento de conocimiento en bases de datos.

1. Introdução

A violência doméstica (VD) contra a mulher é um dos crimes mais desproporcionais e injustos previstos na legislação brasileira. A primeira lei que tratou de tipificar as condutas específicas da VD contra a mulher foi a Lei Maria da Penha. Esta lei cria mecanismos com intuito de prevenir e coibir a violência doméstica e familiar contra à mulher (Brasil, 2006).

A existência de uma lei como a Lei Maria da Penha nem sempre proporciona uma segurança real na vida cotidiana das mulheres, senão, após a sua criação e divulgação tais crimes não deveriam existir. Segundo dados da Secretaria de Segurança Pública do Estado do Paraná (SESP/PR), nos anos de 2018, 2019 e 2020, o sistema de boletim de ocorrência unificado (BOU) registrou em média 70.940 ocorrências de VD contra à mulher maior de 18 anos.

Para tomar decisões que previnam ou coíbam a VD é essencial o conhecimento especializado sobre o tema. A SESP/PR armazena e gerencia todos os dados da segurança pública do estado, para se saber mais sobre a VD, a análise de dados tem um papel fundamental na orientação de gestores na tomada de decisões. Autores da área de gerenciamento estratégico da informação já ressaltavam a importância dos dados nas organizações, "estamos cercados de dados, e as organizações estão repletas de dados que poderiam se tornar informação valiosa para algum usuário diante de um problema decisório" (Mcgee & Prusak, 1994, p. 24).

Quando os dados são numerosos e/ou complexos, as análises tendem a ser demoradas e eventualmente intempestivas. Uma solução possível para encontrar padrões em dados economizando recursos como tempo, é a descoberta de conhecimento em bases de dados (KDD). O KDD se originou da junção de várias áreas de conhecimento, destacando-se estatística, descobrimento de padrões, banco de dados, inteligência computacional e aprendizado de máquina (Goldschmidt & Passos, 2005).

Os dados utilizados nesta pesquisa são secundários, extraídos de duas bases de dados da SESP/PR, limitando-se às vítimas mulheres de VD e feminicídio, maiores de 18 anos, de registros dos anos de 2018, 2019 e 2020.

Tendo ciência de que é comum que bases de dados de VD contenham muitos atributos, é necessário valorar os atributos e determinar de forma objetiva quais são mais relevantes para produzir conhecimento por meio de métodos como o KDD. Observou-se em publicações científicas que dentre as soluções mais empregadas, as técnicas de seleção de atributos e suas diferentes abordagens são relatadas em estudos científicos.

Pesquisas que relacionam KDD e VD é um espaço de pesquisa ainda em evolução no Brasil, fato comprovado pela exiguidade de trabalhos depositados em bases de dados de periódicos, trabalhos acadêmicos, dissertações e teses (Da Silva Costa et al., 2021). Portanto, se objetivou desenvolver um fluxo que possibilitasse a análise preditiva em dados de vítimas fatais e não fatais de VD utilizando estatística descritiva e o KDD.

1.1 Violência doméstica

No Brasil a VD contra à mulher é descrita na Lei Maria da Penha. Segundo essa lei, "configura violência doméstica e familiar contra a mulher qualquer ação ou omissão baseada no gênero que lhe cause morte, lesão, sofrimento físico, sexual ou psicológico e dano moral ou patrimonial" (Brasil, 2006). Em seu 7º artigo, classifica os tipos de violência doméstica em: Física, Psicológica, Sexual, Patrimonial e Moral (Brasil, 2006).

Esta lei supriu a necessidade de tratar a VD de forma mais específica. Anteriormente a esta lei, os casos de VD não tinham uma tipificação de crime específica, então, as condutas criminosas desse tipo de crime eram atribuídas a alguns artigos do Decreto-Lei 2.848 (Brasil, 1940), como o artigo 129 que trata das lesões corporais.

Uma característica que diferencia a VD contra à mulher de outros crimes violentos é a revitimização. Para Manzanares et al., (2011), ocorre a revitimização quando uma vítima sofre a experiência da violência doméstica em diversas ocasiões. Karystianis et al., (2019), analisando dados de VD da Polícia Inglesa, constatou que 45% das vítimas registraram mais de um evento de VD e 27% três ou mais. Tal característica aumenta a complexidade da análise sobre dados dessa natureza.

O feminicídio, qualificador do crime de homicídio, ocorre quando o homicídio é: "contra a mulher por razões da condição de sexo feminino: Considera-se que há razões de condição de sexo feminino quando o crime envolve, violência doméstica e familiar; menosprezo ou discriminação à condição de mulher." (Brasil, 2015).

1.2 Efeitos da violência doméstica

A literatura científica sobre VD tem registrado muitos efeitos negativos desta para com suas vítimas. Em casos de VD em que existe agressão física nos EUA, estudos apontaram como resultados para as vítimas: contusões, lesão de cabeça e pescoço (Karakurt et al., 2016; Liu et al., 2020). Ainda, havendo ou não lesão física, essa tem efeitos negativos sobre a saúde mental de suas vítimas. Distúrbios como estresse pós-traumático, depressão e ansiedade são relatados como consequências negativas da VD (Ko & Kim, 2015; Dias et al., 2022). A pesquisa de Dias et al., (2022), foi realizada com dados da saúde de vítimas de VD de um município do mesmo estado a qual se refere as bases desse estudo, o Paraná. Portanto, acrescenta um outro aspecto que não pode ser observado em dados de segurança pública.

Nos dados da SESP/PR verificou-se a existência de ocorrências com lesão corporal grave, lesão corporal gravíssima, lesão corporal seguida de morte e feminicídio. Estes registros representam os piores desfechos dentro das possibilidades de registro de ocorrências para os casos de VD nesse estado. Mesmo nos casos em que a vítima sobrevive ao evento registrado, os efeitos negativos são permanentes para ela.

1.3 Descoberta de conhecimento em bases de dados

A Descoberta de Conhecimento em Bases de Dados, ou o seu termo original em inglês, Knowledge Discovery in Databases, "tem como objetivo encontrar padrões intrínsecos aos dados nela contidos, apresentando-os de forma a facilitar sua assimilação como conhecimento" (Da Silva et al., 2016).

Dentre os modelos de KDD, o modelo proposto por Fayyad et al., (1996) composto por cinco etapas (Seleção, Pré-processamento, Transformação, Data Mining e Interpretação/Avaliação) se tornou popular na academia. Fato visto quando em revisão de literatura sobre a área, mais da metade das pesquisas relatam sua utilização. Outros modelos também relatados em pesquisas são: SEMMA e o CRISP-DM. Estes foram desenvolvidos com interesse de suprir necessidades de empresas da iniciativa privada (Azevedo & Santos, 2008).

As cinco etapas são altamente interligadas, necessitando de completa realização da primeira para que se possa aplicar a segunda. A etapa de seleção trata da extração de dados alvos da análise. No pré-processamento ocorre limpeza nos dados, cálculos de campos como exemplo, a idade (através da data de nascimento) e seleção de atributos. Na transformação, são feitos procedimentos como a normalizar dados e discretizar atributos numéricos, ou seja, adequar os dados para próxima etapa. No Data Mining são aplicados os algoritmos de mineração de dados que já passaram pelas três etapas anteriores. Nessa última etapa, Interpretação/Avaliação, os pesquisadores interpretam os resultados e os especialistas avaliam a utilidade das descobertas.

A etapa de Data Mining é muito ligada aos objetivos do pesquisador. A tarefa de mineração e o algoritmo utilizado para tal são escolhidos dependendo do que se quer saber sobre os dados. Alguns algoritmos podem produzir resultados mais úteis que outros, como o caso verificado nessa pesquisa, o uso de regras de classificação para descobrir conhecimento sobre bases de dados de VD contra à mulher.

1.4 Regras de classificação

As regras de classificação buscam correspondência entre atributos de bases de dados, geralmente de sistemas especialistas, extraindo conhecimento novo para que especialistas possam interpretar com facilidade (Halmenschlager, 2002).

Uma regra gerada por um algoritmo de indução de regras é mais fácil de se interpretar, haja visa que o resultado é o que mais se aproxima mais da linguagem natural, fato que auxilia a tomada de decisões de gestores baseando-se nas descobertas sobre os dados.

Pesquisando sobre os algoritmos de indução de regras com potencial de uso em pesquisas como as em dados de VD, destacaram-se o PRISM e o CN2.

O PRISM é: "baseado no algoritmo de indução de regras ID3. A motivação na sua criação foi de superar as limitações na representação de árvores de decisão. PRISM produz regras modulares (independentes) que são de mais fácil compreensão." (Vasconcelos, 2002, p. 33).

Por sua vez, o CN2, por ter a sua abordagem de cima para baixo é eficiente em minerar dados ruidosos, além disso, utiliza a entropia para selecionar um nó e determinar sua posição na árvore de decisões. Quando nenhuma informação é significativa o algoritmo interrompe o crescimento da árvore evitando um super ajuste ao modelo dos dados (Clark & Niblett, 1989).

As regras produzidas pelos dois algoritmos são independentes e a lógica de crescimento da árvore é inversa, fato que instiga a curiosidade de observar o seu desempenho quando executados com os mesmos dados.

1.5 Seleção de atributos

Nas pesquisas que utilizam o KDD para criar conhecimento sobre dados é comum que bases de dados sejam extraídas com muitos atributos, determinar quais dos atributos são os mais relevantes tem motivado um número crescente de estudos na academia. Desde o fim da década de 1980, pesquisadores dessa área publicam soluções para reduzir a dimensionalidade dos dados com diversas abordagens. O fato é que, nem todos os atributos contribuem positivamente na etapa de mineração.

As abordagens de seleção de atributos são subdivididas em: Filtro, Wapper, Embedded e hibrida (Lee, 2005). A abordagem filtro é realizada na etapa de pré-processamento e utiliza medida independentes com: dependência, informação, distância, consistência e precisão (Lima, 2016; Macedo, 2012; Souza, 2017). A abordagem Wapper, ou empacotada, busca incessantemente o melhor conjunto de dados através de um algoritmo de mineração de dados usando o desempenho como critério de avalição, porém tem alto custo computacional (Souza, 2017). A abordagem Embedded, ou embutida é realizada no próprio algoritmo de mineração de dados de descobertas de padrões, então, essa não ocorre no pré-processamento como as anteriores (Lima, 2016). Considera-se uma abordagem como híbrida quando as abordagens filtro e wrapper são usadas em conjunto (Liu & Yu, 2005).

Na observação do potencial de cada abordagem e o que agrega na análise, se optou por seguir com a abordagem filtro, com as técnicas: CFS, Info Gain, Gain Ratio e ReliefF. Essas técnicas de seleção de atributos usam medidas independentes e é possível acompanhar a seleção na etapa de pré-processamento, enriquecendo o estudo.

2. Metodologia

A pesquisa foi conduzida com abordagem quantitativa, analisando a frequência de atributos. O método é indutivo, buscando regularidade e padrão nos dados. Segundo Gil (2008, p. 27), as pesquisas exploratórias "têm como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores". Portanto, se buscou trazer um novo conhecimento científico/prático para a área de análise de dados de violência doméstica e feminicídio.

Este tópico trata das bases de dados utilizadas na pesquisa e o percurso metodológico empregado nas análises com estatística descritiva e com KDD para bases de dados de VD com reincidência.

2.1 Bases de dados

A base de dados do sistema BOU disponibilizada pela SESP/PR é composta por 368.476 linhas e 23 atributos (colunas) de ocorrências de VD contra mulheres maiores de idade em 2018, 2019 e 2020. Essa base é resultado da junção das tabelas "boletim" e "vítima". Com exceção de 3 atributos (numerações de sistemas e de identificação da base), cada uma dessas colunas descreve dados do boletim como: data hora, dia da semana, período, cidade, bairro, nome jurídico (nome do crime registrado em lei), idade estimada, estado civil, grau de instrução etc.

A base de dados do SCOL disponibilizada pela SESP/PR é composta por 237 linhas e 28 atributos (colunas) de ocorrências de mortes intencionais catalogadas como feminicídio em 2018, 2019 e 2020. A base de dados conta com atributos como: data, cidade, bairro, autoria conhecida, motivação, idade, raça cor, estado civil etc. Se observou que a base SCOL é composta por mais atributos demográficos que a base BOU. A criação do sistema de ocorrências letais se deu pela necessidade de atualizar dados das ocorrências com mortes, feito pelos policiais civis do estado, já que um boletim de ocorrência encerrado e/ou encaminhado fica bloqueado para alterações.

A terceira e última base de dados, Naturezas, foi solicitada a SESP/PR devido à complexidade de analisar a revitimização utilizando a base BOU. Esta base foi pré-processada no banco de dados do BOU durante a sua extração. Esta é composta por 3 atributos: "chave" (concatenação da chave nome da vítima com o nome da mãe), "ocorrencia" (nome jurídico da conduta criminal) e "count" (contagem da natureza por vítimas). Também, foi limitada a extração para dados de ocorrências de VD contra mulheres maiores de idade em 2018, 2019 e 2020.

2.2 Descoberta de conhecimento em bases de dados

Os diversos procedimentos realizados no KDD dessa análise foram tabulados, organizados por etapa do modelo de Fayyad e são vistos no Quadro 1.

Quadro 1 - Procedimentos da análise com KDD agrupados por etapas.

Etapas do KDD	Software	Procedimentos (Software utilizado)
Seleção	Postgres	Extração dos dados das ocorrências de violência doméstica com vítima mulher de 18 ou mais, dos anos de 2018, 2019 e 2020. Realizado pela SESP/PR.
Pré-processamento	Rstudio, Orange e WEKA	Verificação de completude nos atributos; exclusão de atributos incompletos; alinhamentos de ocorrência por vítima; identificação das vítimas mortas; balanceamento de classe por *uppersampler* e; seleção de atributos com as técnicas: CFS (WEKA), *Info Gain*, *Gain Ratio* e *ReliefF (Orange).*
Transformação	RStudio	Padronização de valores de atributos com conteúdo aberto [Casado/Casada]; simplificação atributos com mais de 40 caracteres e; criação de um campo denominado morte para atribuir a morte de uma vítima na tabela SCOL.
Data mining	Orange, WEKA	Aplicação dos atributos selecionados pelas quatro técnicas de seleção de atributos no algoritmo PRISM (WEKA) e no algoritmo CN2 (*Orange*).
Interpretação/Avaliação	Excel	Interpretação e anotação de todos os resultados dos dois algoritmos; avaliação objetiva por medidas e matriz de confusão e; avaliação subjetiva dos especialistas nas regras aprovadas objetivamente.

Fonte: Autores (2022).

A análise subjetiva foi apoiada no modelo de avaliação de John[1]. As regras são avaliadas em: a) fica satisfeito com os resultados e surpreso com alguns dos padrões obtidos; b) fica satisfeito com os resultados, mas percebe que já conhecia os padrões obtidos; c) fica insatisfeito com os resultados.

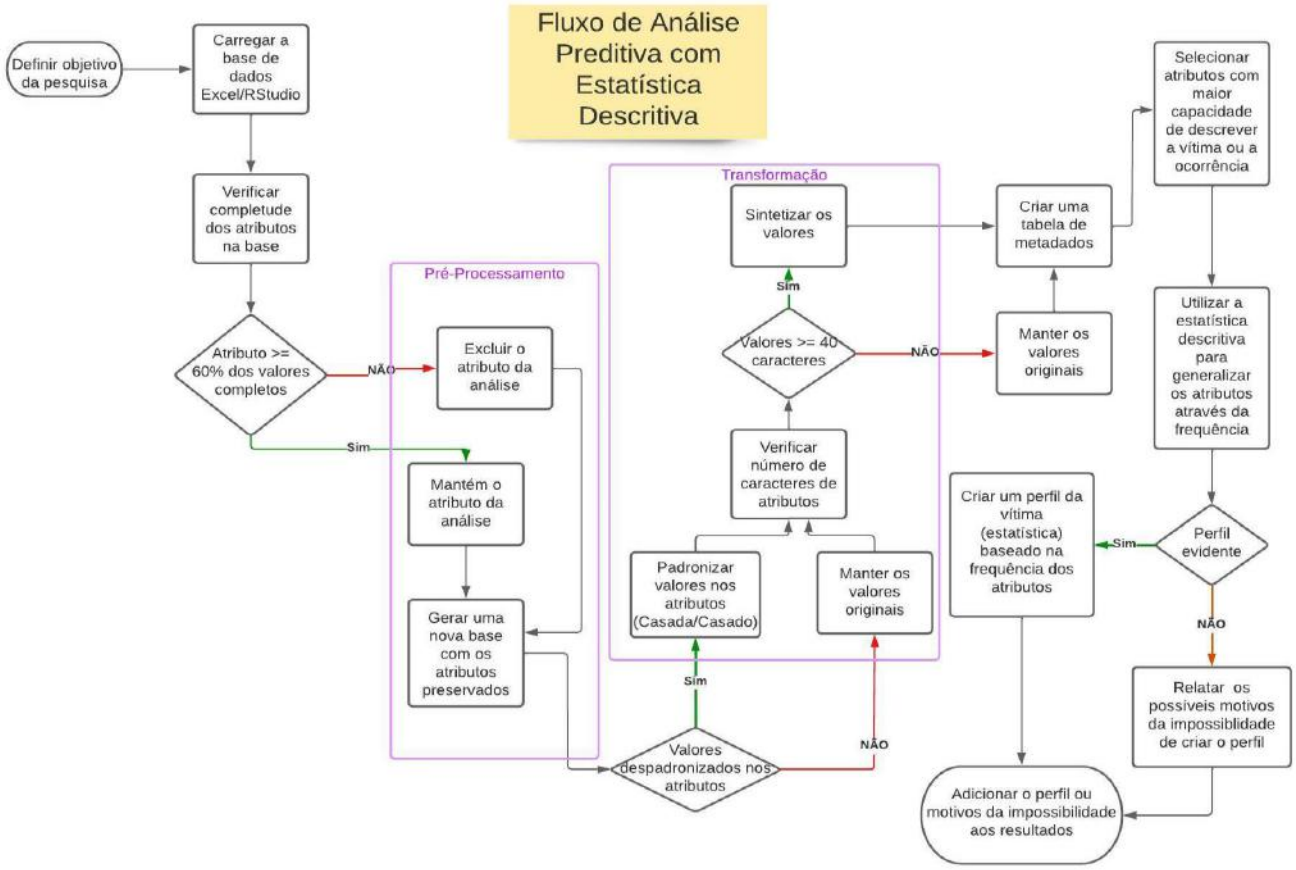
3. Resultados e Discussão

Os resultados da aplicação dos fluxos criados neste estudo foram validados pelos especialistas (analistas criminais). O fluxo de análise preditiva com estatística descritiva e com KDD para bases de dados com reincidências são apresentados abaixo.

O primeiro fluxo descreve os passos necessários para analisar dados com a estatística descritiva de uma base de dados de VD. Os procedimentos estão mais ligados a limpeza da base, exclusão dos atributos de completude abaixo de 60%, padronização de valores e sintetização de valores de atributos com mais de 39 caracteres. Este fluxo apoia pesquisas com objetivos ligados a descrição de base de dados de VD e é indispensável para quem utiliza o fluxo de análise preditiva com KDD em bases com reincidência. O fluxo é visto na Figura 1.

[1] John, G. H. (1997). *Enhacements to the data mining process*. (Doctoral dissertation). Department of Computer Science, Stanford University, Stanford.

Figura 1 – Fluxograma de análise de dados preditiva com estatística descritiva.

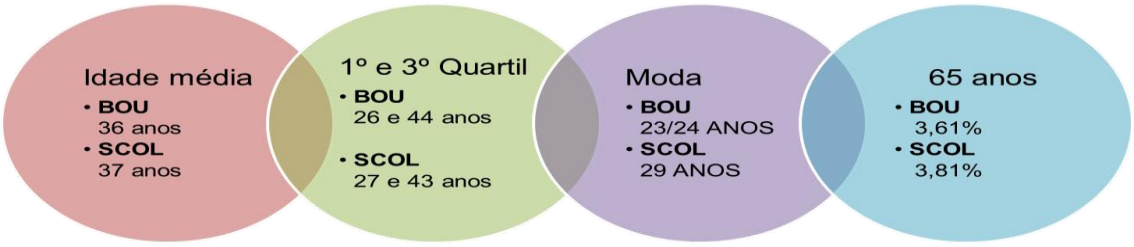


Fonte: Autores (2022).

Através do uso do fluxo de análise preditiva com estatística descritiva foi possível verificar algumas características nas bases, incluindo a revitimização que foi observada em 23.680 (19%) vítimas de violência doméstica na base de dados com reincidência, ainda, 58 vítimas registraram entre 10 e 19 boletins nos três anos estudados.

Alguns atributos com potencial explicativo e que coexistentes nas bases BOU e SCOL nesta análise são comparados. O atributo idade (SCOL) e idade estimada (BOU) demonstraram comportamento parecido em médias e proporções, inclusive para as vítimas de 65 anos ou mais. A moda da base SCOL, 29 anos representou a maior diferença entre as idades registradas nas duas bases de dados. Ressalta-se que em todos os anos foi observado alta dispersão nas duas bases. Tais números são visualizados na Figura 2.

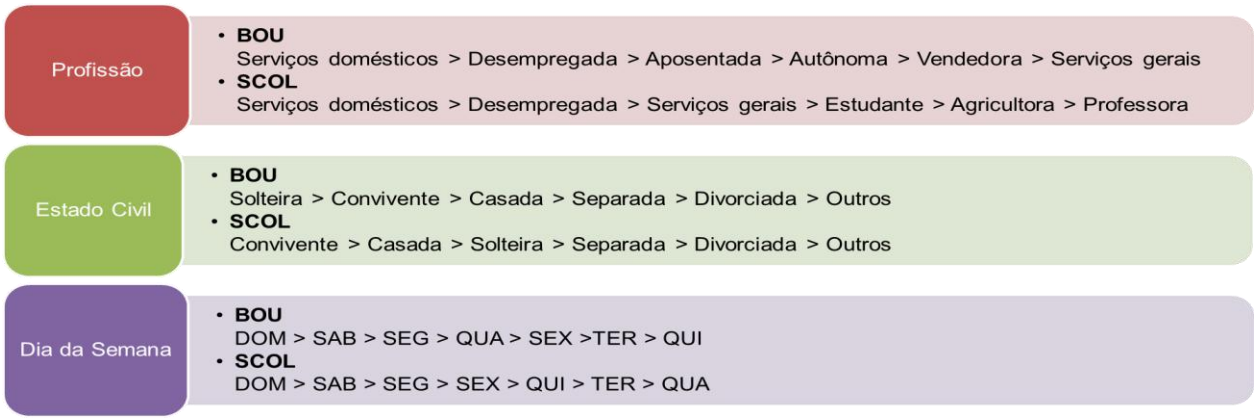
Figura 2 - Estatística descritiva das idades da base BOU *versus* base SCOL.



Fonte: Autores (2022).

Os atributos Profissão, Estado Civil e Dia da Semana não tem potencial de diferenciar significativamente as vítimas das duas bases. Porém, mesmo que os dados estudados tenham um comportamento parecido, é possível verificar que: as vítimas de violência doméstica tem sua ocupação (profissão) relacionados ao serviço doméstico (dentro ou fora de casa) ou desempregada; para a profissão, no BOU se acrescenta Aposentada; o estado civil mais frequente para vítimas do BOU é solteira, seguida de convivente; o estado civil para o SCOL tem uma frequência maior para as conviventes e casadas, então, é mais provável que uma vítima do SCOL coabite com o autor de sua morte; os dias da semana mais frequentes tanto para mortes quanto para ocorrências de VD foram domingo, sábado e segunda. A ordem da frequência dos atributos é verificada na Figura 3.

Figura 3 – Ordem da Frequência de Profissão, Estado Civil e Dia da Semana das bases BOU *versus* SCOL.

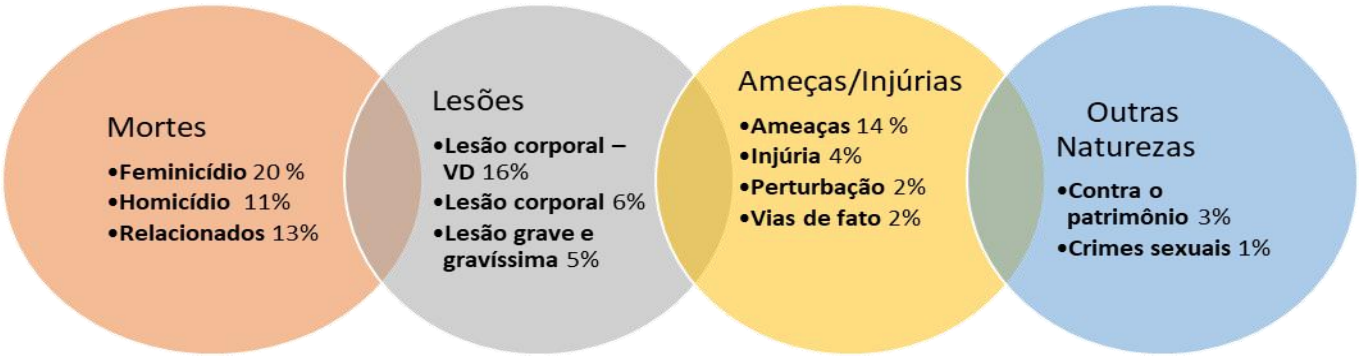


Fonte: Autores (2022).

No intuito de verificar se a ordem da frequência relativa dos dias da semana é modificada com a exclusão de semanas com feriados, constatou-se que a ordem não se altera para os primeiros lugares. Um caso pontual ocorreu com os dados do ano de 2019 para as vítimas do SCOL, as semanas com feriados inflacionam a terça-feira em 7%.

O cruzamento das vítimas das bases de dados SCOL e Naturezas observou que, 149 delas constavam nas duas bases. Então, das 236 vítimas de feminicídio da base SCOL, 149 possuíam histórico de VD registrado no BOU. A frequência relativa das naturezas jurídicas (lesão corporal, ameaça, feminicídio) das 149 vítimas é apresentada na Figura 4.

Figura 4 – Frequência relativa de naturezas jurídicas das vítimas do SCOL com histórico de violência doméstica.



Fonte: Autores (2022).

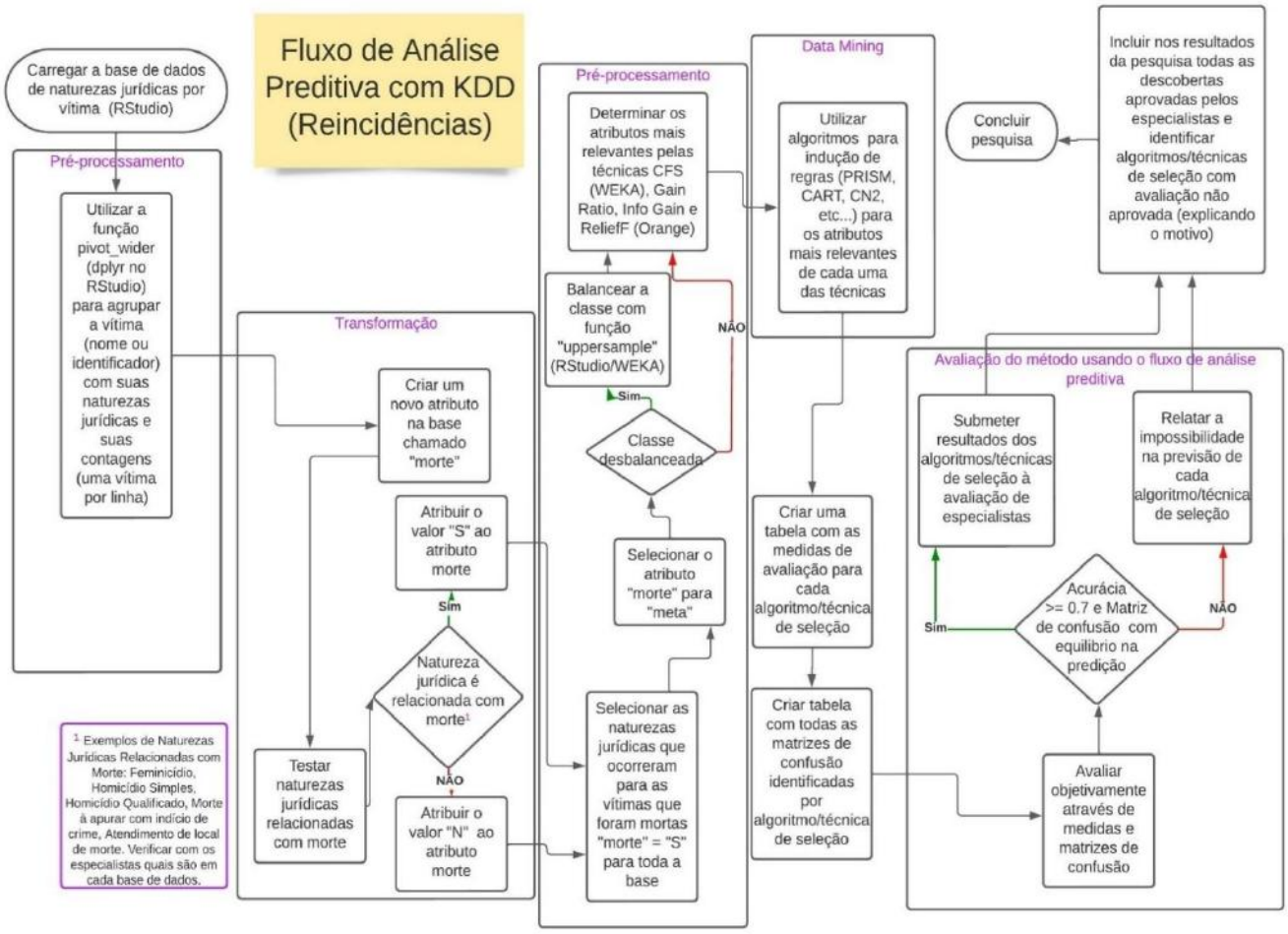
Nota-se que em aproximadamente 43% das ocorrências do BOU para as 149 vítimas, a natureza jurídica cadastrada no boletim era relacionada com morte, como feminicídio, 20%. A qualificação de uma morte como feminicídio no estado do Paraná é realizada pelo delegado de polícia civil. Porém, na investigação, uma morte pode ser requalificada pela própria polícia civil ou por instituições do poder judiciário. Então, a frequência verificada pode ser explicada pela requalificação da natureza jurídica da morte. Assim, na intersecção dos dados dessas duas bases de dados (SCOL e Naturezas) foi constatado que aproximadamente 63% das vítimas de feminicídio tem histórico de VD nos anos estudados.

Por sua vez, o fluxo análise de dados preditiva com KDD em bases de dados com reincidências, descreve os passos de uma análise preditiva em uma base de dados como a Naturezas. Nessa análise foi necessário desdobrar o atributo nome jurídico (lesão corporal, ameaça, vias de fato) e alinhar sua contagem com cada vítima. Esse procedimento, alinhar as vítimas com a contagem dos crimes, realizado nas etapas de pré-processamento e transformação, é necessário quando a base de dados teve um pré-processamento de contagem agrupada, porém o atributo (nome jurídico do crime) se repete nas contagens de algumas vítimas.

Ressalta-se que, foi previsto que devido à legislação vigente no país, pesquisadores não tenham acesso a campos como o nome da vítima, então, na parte inferior esquerda é descrito o procedimento necessário para automatizar o processo de identificação da vítima que foi morta.

Os procedimentos são descritos com maiores detalhes na Figura 5, assim como, a recomendação para identificação das vítimas fatais sem o nome completo da vítima.

Figura 5 – Fluxograma de análise de dados preditiva com KDD em bases com reincidências.



Fonte: Autores (2022).

Os resultados da execução do KDD com o fluxo supracitado nos dados da base Naturezas com vítimas mortas e não mortas foram reprovados na etapa de avaliação objetiva do algoritmo PRISM. As matrizes de confusão apresentaram desequilíbrio, tendo como característica o baixo rendimento em identificar a vítima com anotação de morte. Também, a maior acurácia registrada com os atributos selecionados pelas quatro técnicas de seleção de atributos foi 0,548. Já no algoritmo CN2, a avaliação objetiva foi aprovada. As matrizes de confusão apresentaram equilíbrio para os atributos selecionados pelas quatro técnicas de seleção de atributos. A acurácia acima de 0,710 foi registrada pelas execuções do algoritmo com os atributos selecionados pelas técnicas CFS e Info Gain. Então, as regras geradas por essas técnicas foram enviadas para aprovação subjetiva dos especialistas. Os valores são vistos na Tabela 1.

Tabela 1 - Resultados do algoritmo CN2 com quatro técnicas de seleção de atributos.

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

Fonte: Autores (2022).

Na avaliação subjetiva os especialistas atribuíram insatisfeito para uma regra de cada técnica de seleção aprovada. Os resultados gerados através dos atributos selecionados pela técnica CFS receberam a avaliação mais alta comparada a outra técnica. Uma união de todas as regras aprovadas é demostrada no Quadro 2.

Quadro 2 - Compêndio de regras aprovadas pelos especialistas.

REGRAS RESULTANTES DA PREDIÇÃO DO ALGORITMO CN2 COM AS TÉCNICAS CFS E INFO GAIN (PROBABILIDADE DE MORTE POR REGRA)
Lesão Corporal - Violência Doméstica e Familiar = 1 - 2 (50%), = 3 - 4 (75%) ou = 7 - 8 (67%)
Lesão Corporal - Violência Doméstica e Familiar diferente de 1 - 2 e Ameaça= 0 (74%)
Lesão Corporal - Violência Doméstica e Familiar diferente de 1 - 2 e Ameaça diferente de 1 - 2 (74%)
Lesão Corporal - Violência Doméstica e Familiar = 0 (50%)
Lesão Corporal de Natureza Grave = 1 - 2 (83%)
Lesão Corporal de Natureza Gravíssima = 1 - 2 (90%)
Descumprir Decisão Judicial que Defere Medidas Protetivas de Urgência Previstas nesta Lei = 1 - 2 e Lesão Corporal - Violência Doméstica e Familiar diferente de 0 (67%)
Descumprir Decisão Judicial que Defere Medidas Protetivas de Urgência Previstas nesta Lei = 1 - 2 e AMEACA diferente de 1 - 2 (75%)
Ameaça = 0 e boletins diferente de 1 - 2 (83%)
Constrangimento Ilegal = 1 - 2 (67%)
Vias de Fato = 1 - 2 e boletins = 1 - 2 (71%)

Fonte: Autores (2022).

O Quadro 2 demonstra através das regras induzidas as probabilidades de morte para cada vítima por regra. Como exemplo, uma vítima dessa base que tem o registro de uma a duas ocorrências com natureza “vias de fato” e de um a dois registros na contagem geral de boletins, essa vítima tem 71% de morte. Regra vista na última linha do Quadro 2.

4. Considerações Finais

As análises nos dados de VD do estado do Paraná evidenciaram uma realidade ainda triste sobre a violência sofrida pelas mulheres maiores de 18 anos. Mesmo que, tanto o número de ocorrências quanto o número de mortes apresentam queda em 2020, ainda estamos distantes de uma sociedade mais igual, que não trata a mulher com violência.

O resultado das análises de dados preditivas com KDD em base com reincidências foram aprovadas e algumas regras surpreenderam positivamente os especialistas. Ainda, esses especialistas relataram que conhecimentos evidenciados pelas regras induzidas, foram adquiridos pós muitas horas de análise de dados acessando mais de dois sistemas diferentes. Então, a análise preditiva realizada nessa pesquisa contribui com a economia de tempo nas análises feitas pelos especialistas.

Os fluxos desenvolvidos nesta pesquisa são interligados. Portanto, para que se possa aplicar o fluxo de análise preditiva com KDD é necessário a realização dos passos do fluxo de análise com estatística descritiva. A ideia inicial era criar um único fluxo de análise preditiva, porém, separá-los auxilia na assimilação e apoia pesquisadores com objetivos diferentes. Pesquisas com o objetivo de criar perfil de vítimas de VD através da estatística descritiva são auxiliadas pelo fluxo da Figura 1. Para pesquisadores que pretendem criar perfil de vítimas através do KDD em bases com reincidências, o fluxo da Figura 5 descreve todos os procedimentos a serem adotados por estes.

Ressalta-se que a análise de dados de violência doméstica, principalmente em dados com reincidência não é uma tarefa trivial de análise com KDD. Este tipo de análise requer um esforço maior nas etapas de pré-processamento e transformação, o que pode muitas vezes desencorajar pesquisadores a analisar dados de VD. Então, os fluxos descritos apoiam os pesquisadores com interesse de pesquisar VD e feminicídio, oferecendo um roteiro que já produziu resultados aprovados por analistas especializados nos temas. Esperasse que esse apoio oferecido pelos fluxos incentive novas pesquisas sobre temas de relevância social e acadêmica, como os tratados aqui.

Propõe-se para estudos futuros a mineração de processos na base BOU e exploração da capacidade de previsão das redes neurais na base Naturezas. O resultado da mineração de processos apoiaria decisões que podem apontar para a atualização da infraestrutura da base BOU ou no processo de extração de dados, tudo no sentido de auxiliar a descoberta em processos de KDD e de outras análises de dados realizadas pela SESP/PR.

Referências

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In A. Abraham (ed.), *IADIS European Conference Data Mining*, IADIS, 182-185. http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf

Brasil. (31 de Dezembro de 1940). Decreto-Lei 2.848, de 07 de Dezembro de 1940. *Legislação*. Brasília, Distrito Federal, Brasil.

Brasil. (7 de Agosto de 2006). Lei nº 11.340, de 7 de Agosto de 2006. *Legislação*. Brasília, Distrito Federal, Brasil.

Brasil. (9 de Março de 2015). Lei nº 13.104, de 9 de Março de 2015. *Legislação*. Brasília, Distrito Federal, Brasil.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm machine learning. *Machine Learning*, [s. l.], 3, 261-283. https://link.springer.com/content/pdf/10.1023/A:1022641700528.pdf

Da Silva Costa, C. A., Tsunoda, D. F., & Pecini, A. C. (2021). Análise de dados de violência doméstica: revisão integrativa. In: *XXVI Congresso Nacional de Administração*, Goiânia: SINAGO. https://drive.google.com/file/d/1R-y8s_mEP6LrelU9NW2OQq2jnoSWTelL/view

Da Silva, L. A., Peres, S. M., & Boscarioli, C. (2016). *Introdução à mineração de dados: com aplicação em R.* (1 ed.). Rio de Janeiro: Elsevier.

Dias, E. R., Uscocovich, K. J. S., & Lise, A. M. R. (2022). Post-traumatic stress disorder in women who suffer domestic violence in the city of Cascavel-PR. *Research, Society and Development*, [S. l.], 11, (17), e101111738850. doi:https://doi.org/10.33448/rsd-v11i17.38850

Fayyad, U., Piatestky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining. *Association for the Advancement of Artificial Intelligence* (AAAI), [s. l.], 17, (3), 37–54. https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf

Gil, A. C. (2008). *Métodos e técnicas de pesquisa social*. (6 ed.). São Paulo: Atlas.

Goldshmidt, R., & Passos, E. (2005). *Data mining: conceitos, técnicas, algoritmos, orientações e aplicações.* Rio de Janeiro: Elsevier.

Halmenschlager, C. (2002). *Um algoritmo para indução de árvores e regras de decisão*. (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul, Porto Alegre. https://www.lume.ufrgs.br/bitstream/handle/10183/2755/000325797.pdf

Karakurt, G., Patel, V., Whiting, K., & Koyutürk, M. (2016). Mining electronic health records data: domestic violence and adverse health effects. *Journal of Family Violence*, 32, (1), 79–87. doi:https://doi.org/10.1007/s10896-016-9872-5

Karystianis, G., Adily, A., Schofield, P. W., Greenberg, D., Jorm, L., Nenadic, G., & Butler, T. (2019). Automated analysis of domestic violence police reports to explore abuse types and victim injuries: text mining study. *Journal of Medical Internet Research*, 21, (3), e13067. doi:https://doi.org/10.2196/13067

Ko, K. S., & Kim, M. S. (2015). Grounded theory approach on post-divorce social adjustment experience of female victims of domestic violence. *Indian Journal of Science and Technology*, 8, (18). doi:https://doi.org/10.17485/ijst/2015/v8i18/77013

Lee, H. D. (2005). *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. (Tese de Doutorado). Universidade de São Paulo, São Carlos. https://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/publico/tese_huei.pdf.

Lima, R. A. F. de. (2016). *Estratégias de seleção de atributos para detecção de anomalias em transações eletrônicas*. (Dissertação de Mestrado). Universidade Federal de Minas Gerais, Belo Horizonte. https://www.dcc.ufmg.br/pos/cursos/defesas/1930M.PDF

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, [s. l.], 17, (4), 491–502. https://www.cs.binghamton.edu/~lyu/publications/Liu-Yu05TKDE.pdf

Liu, L. Y., Bush, W. S., Koyutürk, M., & Karakurt, G. (2020). Interplay between traumatic brain injury and intimate partner violence: Data Driven Analysis utilizing electronic health records. *BMC Women's Health*, 20, (1). doi:https://doi.org/10.1186/s12905-020-01104-4

Macedo, D. C. de. (2012). *Comparação da redução de dimensionalidade de dados usando seleção de atributos e conceito de framework: um experimento no domínio de clientes*. (Dissertação de Mestrado). Universidade Tecnológica Federal do Paraná, Ponta Grossa. https://repositorio.utfpr.edu.br/jspui/bitstream/1/602/3/PG_PPGEP_M_Macedo%2C%20Dayana%20Carla%20de_2012.pdf

Manzanares, R. C., Tarrío, C. T., & Salgado, C. A. (2011). Mediacón em violencia de género. *Revista de Mediación*, [s. l.], 4, (7), 38-45. https://revistademediacion.com/wp-content/uploads/2013/10/Revista-Mediacion-7-05.pdf

Mcgee, J. V., & Prusak, L. (1994). *Gerenciamento estratégico da informação: aumente a competitividade e a eficácia de sua empresa utilizando a informação como uma ferramenta estratégica*. Rio de Janeiro: Campus.

Souza, J. T. de. (2017*). Métodos de seleção de atributos e análise de componentes principais: um estudo comparativo*. (Dissertação de Mestrado). Universidade Tecnológica Federal do Paraná, Ponta Grossa. https://repositorio.utfpr.edu.br/jspui/bitstream/1/2387/1/PG_PPGEP_M_Souza%2C%20Jovani%20Taveira%20de_2017.pdf

Vasconcelos, B. de S. (2002). *Mineração de regras de classificação com sistemas de banco de dados objeto-relacional: estudo de caso: classificação de litofácies de poços de petróleo*. (Dissertação de Mestrado). Universidade Federal de Campina Grande, Campina Grande. http://docs.computacao.ufcg.edu.br/posgraduacao/dissertacoes/2002/Dissertacao_BenitzDeSouzaVasconcelos.pdf