

---

# **Técnicas de Agrupamento aplicadas aos Dados de Acidente de Trabalho**

---

**Daniela Freitas Giacomelli**



UNIVERSIDADE FEDERAL DE UBERLÂNDIA  
FACULDADE DE COMPUTAÇÃO  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia  
2020

**Daniela Freitas Giacomelli**

**Técnicas de Agrupamento aplicadas aos Dados  
de Acidente de Trabalho**

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof<sup>a</sup>. Dr<sup>a</sup>. Elaine Ribeiro de Faria Paiva  
Coorientador: Prof. Dr. Murilo Coelho Naldi

Uberlândia  
2020

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU  
com dados informados pelo(a) próprio(a) autor(a).

G429      Giacomelli, Daniela Freitas, 1991-  
2020      Técnicas de agrupamento aplicadas aos dados de acidentes de  
trabalho [recurso eletrônico] / Daniela Freitas Giacomelli. - 2020.

Orientadora: Elaine Ribeiro de Faria Paiva.  
Coorientador: Murilho Coelho Naldi.  
Dissertação (Mestrado) - Universidade Federal de Uberlândia,  
Pós-graduação em Ciência da Computação.  
Modo de acesso: Internet.  
Disponível em: <http://doi.org/10.14393/ufu.di.2020.487>  
Inclui bibliografia.  
Inclui ilustrações.

1. Computação. I. Paiva, Elaine Ribeiro de Faria, 1980-, (Orient.).  
II. Naldi, Murilho Coelho, 1981-, (Coorient.). III. Universidade  
Federal de Uberlândia. Pós-graduação em Ciência da Computação.  
IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:  
Gizele Cristine Nunes do Couto - CRB6/2091  
Nelson Marcos Ferreira - CRB6/3074


**UNIVERSIDADE FEDERAL DE UBERLÂNDIA**

Coordenação do Programa de Pós-Graduação em Ciência da Computação  
 Av. João Naves de Ávila, nº 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902  
 Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br


**ATA DE DEFESA - PÓS-GRADUAÇÃO**

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Mestrado Acadêmico, 18/2020, PPGCO				
Data:	22 de junho de 2020	Hora de início:	14:07	Hora de encerramento:	16:40
Matrícula do Discente:	11812CCP014				
Nome do Discente:	Daniela Freitas Giacomelli				
Título do Trabalho:	Técnicas de agrupamento aplicadas aos dados de acidentes de trabalho				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se, por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Humberto Luiz Razente - FACOM/UFU; Ricardo Cerri - CCET/UFSCAR; Murilo Coelho Naldi - CCET/UFSCAR (coorientador) e Elaine Ribeiro de Faria Paiva - FACOM/UFU, orientadora da candidata.

Os examinadores participaram desde as seguintes localidades: Humberto Luiz Razente - Glendale, Califórnia, Estados Unidos da América; Ricardo Cerri e Murilo Coelho Naldi - São Carlos - SP; Elaine Ribeiro de Faria Paiva - Uberlândia-MG. A discente participou da cidade de Uberlândia-MG.

Iniciando os trabalhos a presidente da mesa, Profa. Dra. Elaine Ribeiro de Faria Paiva, apresentou a Comissão Examinadora e a candidata, agradeceu a presença do público, e concedeu à Discente a palavra para a exposição do seu trabalho. A duração da apresentação da Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir a senhora presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir a candidata. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando a candidata:

Aprovada.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.



Documento assinado eletronicamente por **Ricardo Cerri, Usuário Externo**, em 22/06/2020, às 18:49, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Humberto Luiz Razente, Professor(a) do Magistério Superior**, em 22/06/2020, às 19:25, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Elaine Ribeiro de Faria Paiva, Professor(a) do Magistério Superior**, em 23/06/2020, às 09:18, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Murilo Coelho Naldi, Usuário Externo**, em 23/06/2020, às 10:37, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site [https://www.sei.ufu.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **2095786** e o código CRC **AFFC4D01**.

---

# Agradecimentos

Meus agradecimentos aos familiares e amigos pelo apoio ao longo do desenvolvimento desta pesquisa. Em especial, aos meus pais, Everson e Maria, que nunca mediram esforços para me proporcionar as melhores condições de estudo. Ao meu irmão, Everson Junior, e à minha tia, Elaine Giacomelli, que compartilharam comigo as angústias e nunca me deixaram desistir.

Ao meu marido Fernando, por todo amor e compreensão. Meu companheiro, que me incentiva e acredita em mim, mesmo quando eu não acredito. Se agora encerro esse ciclo, devo isso a você. À minha querida Manuela, fruto desse relacionamento, que ainda nem nasceu, mas já é um incentivo para eu buscar sempre a minha melhor versão.

À minha orientadora Elaine, por ter me aceitado como sua aluna de mestrado. Agradeço pela dedicação em realizar leituras tão minuciosas e fazer suas considerações, pela disponibilidade para reunir quantas vezes fossem necessárias e prontidão em esclarecer minhas dúvidas.

Aos meus colegas Luiz Eugênio e Bruna Alves, que foram de fundamental importância para o desenvolvimento e suporte para execução dos experimentos realizados.

Ao Murilo, meu coorientador, pelas leituras e contribuições com este trabalho. Aos membros da banca, por terem aceitado o convite. Ao professor João Gama e aos professores do Programa de Pós-Graduação em Ciência da Computação da UFU, que tanto contribuíram para a minha formação acadêmica.

À FAPEMIG pela bolsa. Ao Ministério Público do Trabalho, pelo fornecimento dos dados analisados nesta pesquisa.

A Deus, pela vida e por todas as oportunidades a mim concedidas.

---

## Resumo

O Brasil ocupa a 4ª posição no *ranking* mundial de acidentes de trabalho catalogados. Dentre outros infortúnios, tais ocorrências geram transtornos aos acidentados, perdas na produtividade laboral e pressionam o orçamento público referente aos auxílios e indenizações acidentárias. Esta dissertação objetiva buscar e caracterizar grupos de acidentes de trabalho, conferindo interpretabilidade aos resultados obtidos, a fim de extrair informações que possam ser relevantes aos gestores públicos. Para tanto, os procedimentos metodológicos perpassam pela efetivação de um conjunto de etapas, a saber: pré-processamento dos dados; criação de subconjuntos da base original; seleção dos melhores atributos para realizar a tarefa de agrupamento; aplicação de dois algoritmos de agrupamento hierárquicos, HDBSCAN\* e COBWEB; avaliação dos resultados por meio do uso da medida de validação Silhueta Simplificada e emprego da ferramenta PowerBI, para visualizar gráficos que possibilitem avaliar a composição dos grupos encontrados. Tendo isso em vista, fez-se necessária a proposição de uma medida para calcular distância entre duas instâncias, compostas tanto por atributos numéricos como por categóricos. Essa medida possibilitou, na base de dados do presente estudo, a execução de algoritmos relacionais, como o HDBSCAN\*, além do cálculo de medidas de validação que mensura a distância entre instâncias, como a Silhueta Simplificada. Os resultados indicam que a medida de distância aqui proposta dificultou a busca de grupos pelo algoritmo. Dessa forma, para certos casos, nenhum grupo foi encontrado, e, para outros, o algoritmo agrupou somente instâncias idênticas. Não apresentando tais inconvenientes, o algoritmo Cobweb não demandou adaptações para trabalhar com os tipos de dados presentes na base, sendo capaz de agregar não apenas as instâncias idênticas, como também as instâncias similares. A pesquisa evidenciou a susceptibilidade dos trabalhadores do sexo masculino, com idade entre 18 e 34 anos, aos acidentes de trabalho que ocasionam lesões nos dedos das mãos, pelo manuseio de máquinas e equipamentos e/ou ferramentas manuais, sobretudo os que exercem as atividades de Pesca e Aquicultura. As ocorrências dessa natureza ganharam destaque, tanto nos maiores grupos de cada ano como no Triângulo Mineiro/Alto Paranaíba e Metropolitana de São Paulo, as duas mesorregiões analisadas. Não obstante, os grupos

compostos majoritariamente por vítimas do sexo feminino possuem um delineamento um pouco diferente, com destaque àquelas que atuam na fabricação de celulose, papel e produtos correlatos. Ainda que os dedos das mãos continuem como a parte do corpo mais afetada, as trabalhadoras deste segmento estão suscetíveis a incidentes ocasionados pelo manejo de agentes químicos, biológicos e/ou ferramentas manuais.

**Palavras-chave:** Acidentes de Trabalho. Agrupamento de dados. Descoberta de Conhecimento em Bases de Dados. Comunicações de Acidentes de Trabalho.



---

# Abstract

Brazil occupies the fourth place in the worldwide ranking of catalogued labor accidents. Among other misfortunes, such occurrences generate inconveniences to the injured ones, losses in work productivity and pressure the public budget referring to aids and indemnities due to accidents. This dissertation aims to search and characterize groups of labor accidents, granting interpretability to the obtained results, in order to extract information that can be relevant to public managers. Therefore, the methodological procedures go by the implementation of a set of steps, namely: data pre-processing; creation of subsets from the original dataset; selection of the best attributes to the clustering task; application of two hierarchical clustering, HDBSCAN\* and COBWEB; evaluation of the results through the use of the Simplified Silhouette validation measure and the use of the PowerBI tool, to visualize graphics which may able the evaluation and the composition of the clusters found. Therefore, it was necessary to propose a measure to calculate the distance between two instances, composed as by numerical attributes as by categorical ones. This measure enabled, in the dataset of the present study, the execution of relational algorithms, such as HDBSCAN\*, besides the calculus of validation measures which measures the distance between instances, such as the Simplified Silhouette. The results show that the distance measure here proposed made the search of clusters by the algorithm hard. Thus, to certain cases, no clusters were found, and, to the other ones, the algorithm clustered only identical instances. Not presenting such inconvenient, the Cobweb algorithm didn't demand adaptations to work with the kind of data present in the basis, being able to aggregate not only identical instances, but also similar instances. The research demonstrated the susceptibility of male workers, with the age between 18 and 34 years old, the labor accidents which cause injures on the fingers, by handling machines and equipment and/or manual tools, moreover the ones who perform activities such as Fishing and Fish Farming. The occurrences of this nature gained prominence, such as in bigger clusters of each year as in Triângulo Mineiro/Alto Paranaíba and Metropolitan São Paulo, both analyzed mesoregions. Nevertheless, the clusters composed mostly by female victims have a slightly different delineation, especially those who work in the production of cellulose,

paper and correlated products. Even though the fingers continue as the most affected body part, the female workers of this segment are likely to accidents occasioned by the management of chemical agents, biological ones and/or manual tools.

**Keywords:** Labor accidents. Clustering of data. Knowledge Discovery in Database. Communication of labor accidents..

---

## Lista de ilustrações

Figura 1 – Etapas do processo de KDD. . . . .	24
Figura 2 – Dendrograma obtido pelo método de agrupamento hierárquico. . . . .	30
Figura 3 – Árvore de grupos - HDBSCAN* . . . . .	36
Figura 4 – Exemplo de uma árvore de classificação produzida pelo COBWEB . . .	37
Figura 5 – Etapas do método proposto . . . . .	57
Figura 6 – Hierarquia HDBSCAN* . . . . .	69
Figura 7 – Hierarquia HDBSCAN* - Ano 2017 . . . . .	70
Figura 8 – Grupos 2017 - corte acima do FOSC . . . . .	70

---

## Lista de tabelas

Tabela 1 – Codificação inteira-binária . . . . .	28
Tabela 2 – Codificação 1-de-n . . . . .	29
Tabela 3 – Trabalhos relacionados . . . . .	45
Tabela 4 – Base de Dados de acidentes de trabalho . . . . .	56
Tabela 6 – Parametrização HDBSCAN* . . . . .	66
Tabela 7 – Estudo dos atributos . . . . .	68
Tabela 8 – Maior grupo de cada ano (HDBSCAN*) . . . . .	72
Tabela 9 – Maiores grupos - Mesorregião 3515 (HDBSCAN*) . . . . .	75
Tabela 10 – Maiores grupos - Mesorregião 3105 (HDBSCAN*) . . . . .	76
Tabela 11 – Grupos com diferentes partes do corpo atingidas . . . . .	78
Tabela 12 – Comparação de hierarquias usando HAI . . . . .	82
Tabela 13 – Coeficiente de Silhueta Simplificada (Cobweb) . . . . .	83
Tabela 14 – Maior grupo de cada ano (Cobweb*) . . . . .	83
Tabela 15 – Maiores grupos - Mesorregião 3515 (Cobweb) . . . . .	85
Tabela 16 – Maiores grupos - Mesorregião 3105 (Cobweb) . . . . .	86
Tabela 17 – Grupos com diferentes partes do corpo atingidas . . . . .	88
Tabela 18 – Comparação de agrupamentos - HDBSCAN* x Cobweb . . . . .	89

---

# Sumário

<b>1</b>	<b>INTRODUÇÃO . . . . .</b>	<b>15</b>
<b>1.1</b>	<b>Motivação . . . . .</b>	<b>17</b>
<b>1.2</b>	<b>Objetivos e Desafios da Pesquisa . . . . .</b>	<b>17</b>
<b>1.3</b>	<b>Hipóteses . . . . .</b>	<b>18</b>
<b>1.4</b>	<b>Contribuições . . . . .</b>	<b>18</b>
<b>1.5</b>	<b>Organização do Trabalho . . . . .</b>	<b>19</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA . . . . .</b>	<b>20</b>
<b>2.1</b>	<b>Comunicação de Acidente de Trabalho - CAT . . . . .</b>	<b>20</b>
<b>2.2</b>	<b>Medidas do MPT para prevenir acidentes . . . . .</b>	<b>21</b>
2.2.1	Observatório Digital de Saúde e Segurança do Trabalho . . . . .	22
<b>2.3</b>	<b>Descoberta de Conhecimento em Bases de Dados . . . . .</b>	<b>23</b>
2.3.1	Seleção dos dados . . . . .	24
2.3.2	Pré-processamento . . . . .	24
2.3.3	Transformação . . . . .	25
2.3.4	Mineração de Dados . . . . .	25
2.3.5	Interpretação . . . . .	26
<b>2.4</b>	<b>Agrupamento . . . . .</b>	<b>26</b>
2.4.1	Medidas de proximidade . . . . .	27
2.4.2	Métodos particionais . . . . .	29
2.4.3	Métodos hierárquicos . . . . .	30
2.4.4	Métodos baseados em densidade . . . . .	31
<b>2.5</b>	<b>HDBSCAN* . . . . .</b>	<b>32</b>
2.5.1	Definições do DBSCAN* . . . . .	33
2.5.2	Definições HDBSCAN* . . . . .	33
2.5.3	O Algoritmo HDBSCAN* . . . . .	34
2.5.4	Agrupamento não-hierárquico ótimo . . . . .	35
<b>2.6</b>	<b>COBWEB . . . . .</b>	<b>37</b>

2.7	Comparação de hierarquias . . . . .	39
2.8	Medidas de validação . . . . .	40
2.9	Considerações Finais . . . . .	42
3	TRABALHOS RELACIONADOS . . . . .	43
3.1	Abordagem das ciências humanas . . . . .	43
3.2	Abordagem usando técnicas de extração de conhecimento . . .	44
3.3	Considerações finais . . . . .	51
4	MÉTODO PARA DETECÇÃO E ANÁLISE DE GRUPOS NA CATWEB . . . . .	52
4.1	Base de dados . . . . .	52
4.2	Método proposto . . . . .	57
4.2.1	Etapa 1 - Pré-processamento . . . . .	57
4.2.2	Etapa 2 - Criação de subconjuntos . . . . .	59
4.2.3	Etapa 3 - Seleção de atributos . . . . .	59
4.2.4	Etapa 4 - Algoritmos de agrupamento . . . . .	60
4.2.5	Etapa 5 - Medida de validação . . . . .	61
4.2.6	Etapa 6 - Verificação dos resultados . . . . .	62
4.2.7	Etapa 7 - Avaliação do especialista . . . . .	62
4.3	Adaptações e estratégias empregadas para uso dos algoritmos de agrupamento . . . . .	63
4.3.1	HDBSCAN* . . . . .	63
4.3.2	Cobweb . . . . .	63
4.4	Considerações Finais . . . . .	64
5	EXPERIMENTOS E ANÁLISE DOS RESULTADOS - HDBS- CAN* . . . . .	65
5.1	Divisão da base em subconjuntos de dados . . . . .	65
5.2	Parametrização do algoritmo HDBSCAN* . . . . .	66
5.3	Estudo dos atributos . . . . .	67
5.4	Análise do corte da hierarquia . . . . .	69
5.5	Resultados do agrupamento usando o HDBSCAN* . . . . .	71
5.5.1	Experimento 1: . . . . .	71
5.5.2	Experimento 2: . . . . .	73
5.5.3	Experimento 3: . . . . .	74
5.5.4	Experimento 4: . . . . .	77
5.6	Considerações finais . . . . .	78

<b>6</b>	<b>EXPERIMENTOS E ANÁLISE DOS RESULTADOS - COBWEB</b>	<b>80</b>
<b>6.1</b>	<b>Diretrizes para condução dos experimentos . . . . .</b>	<b>80</b>
<b>6.2</b>	<b>Parametrização do algoritmo Cobweb . . . . .</b>	<b>81</b>
<b>6.3</b>	<b>Comparação entre hierarquias . . . . .</b>	<b>81</b>
<b>6.4</b>	<b>Resultados do agrupamento usando o Cobweb . . . . .</b>	<b>82</b>
6.4.1	Experimento 5: . . . . .	82
6.4.2	Experimento 6: . . . . .	84
6.4.3	Experimento 7: . . . . .	87
<b>6.5</b>	<b>Comparação de resultados - HDBSCAN* x Cobweb . . . . .</b>	<b>88</b>
<b>6.6</b>	<b>Considerações Finais . . . . .</b>	<b>89</b>
<b>7</b>	<b>CONCLUSÃO . . . . .</b>	<b>91</b>
<b>7.1</b>	<b>Principais Contribuições . . . . .</b>	<b>92</b>
<b>7.2</b>	<b>Trabalhos Futuros . . . . .</b>	<b>93</b>
	<b>REFERÊNCIAS . . . . .</b>	<b>94</b>

---

## Introdução

Subordinado ao Ministério Público da União (MPU), o Ministério Público do Trabalho (MPT) é o órgão responsável por mediar o vínculo entre empregadores e empregados, supervisionando as relações de trabalho e fiscalizando o cumprimento das normas trabalhistas. Para tanto, respalda-se nas normas da Consolidação das Leis Trabalhistas (CLT).

No início de 2018, segundo levantamento realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), o número de trabalhadores formais totalizou 33,3 milhões de pessoas (IBGE, 2018). No entanto, um elemento que gera preocupação é o fato de que o Brasil se destaca por ser um dos países com a maior ocorrência de acidentes de trabalho (MENDONCA, 2017).

De acordo com o Anuário Estatístico de Acidentes de Trabalho, disponibilizado pela Previdência Social, são considerados acidentes de trabalho:

- ❑ aqueles que ocorrem no exercício do trabalho;
- ❑ aqueles que ocorrem no trajeto entre a residência e o local de trabalho;
- ❑ a doença profissional decorrente do exercício do trabalho ou que com ele se relacione diretamente;
- ❑ o acidente relacionado ao trabalho, mesmo que não tenha sido causa única;
- ❑ agressão, sabotagem ou terrorismo, sofridos no local e horário de trabalho, causados por terceiros ou colegas.
- ❑ desastres como inundação, incêndio ou outros casos decorrentes de força maior;
- ❑ doença proveniente de contaminação do empregado no exercício de sua atividade;
- ❑ acidentes sofridos quando o segurado estiver executando tarefas ou serviços à empresa, mesmo que não seja no local e horário de trabalho.



Destarte, é necessário que os acidentes de trabalho sejam constatados por meio de perícia médica, realizada pelo Instituto Nacional do Seguro Social (INSS). Além de estabelecer a relação entre o acidente e o trabalho, o médico-perito é o profissional responsável por decidir sobre o afastamento do segurado ou o seu retorno imediato às atividades laborais. Nessa condição, o registro de um acidente de trabalho, de trajeto ou de uma doença ocupacional, é feito por meio de um documento denominado Comunicação de Acidente de Trabalho (CAT).

Por sua vez, a empresa deve emitir um CAT até o primeiro dia útil seguinte ao dia do acidente. No caso de morte, a comunicação deve ser imediata. O descumprimento desses prazos acarreta em multa ao empregador. Todavia, se a empresa não confeccionar o mencionado documento, podem fazê-lo o trabalhador, o dependente, a entidade sindical, o médico perito ou a autoridade pública responsável (INSS, 2018).

O Ministério Público do Trabalho, em parceria com a Organização Internacional do Trabalho, lançou o Observatório Digital de Saúde e Segurança do Trabalho (OIT, 2017). Trata-se de uma plataforma digital que fornece visões sobre as CAT's registradas, gastos previdenciários, mortes acidentárias e perfil das vítimas, dentre outras informações. Embora seja uma iniciativa interessante, o Observatório Digital ainda encontra algumas barreiras, como o vasto volume de dados, que está em constante atualização. Este fator constitui uma dificuldade para minerar os dados e extrair informações relevantes.

Com efeito, a análise acurada do conteúdo das CAT's pode oferecer, aos gestores públicos: i) evidências acerca da natureza dos acidentes de trabalho mais frequentes; ii) o estabelecimento de potenciais relações entre região geográfica e os tipos de acidentes recorrentes; iii) a viabilidade de conduzir políticas públicas em locais diferentes que apresentem o mesmo delineamento; iv) a detecção de padrões que poderiam, de outra forma, permanecer ignorados; v) o direcionamento de pesquisas e procedimentos que ambicionem a prevenção de acidentes, de doenças ocupacionais, e o controle dos gastos com benefícios acidentários.

Diante desse cenário, a Mineração de Dados (MD) sobressai como uma possível solução para o problema, dada a sua capacidade de obter informações úteis a partir de grandes depósitos de dados. Ademais, trata-se de uma etapa intrínseca do processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Database* - KDD), que consiste em aplicar algoritmos específicos para extrair padrões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Dentre as diferentes tarefas da Mineração de Dados, é possível destacar a classificação, a regressão, o agrupamento e a associação. Tal como destaca Silva (2018), a escolha da tarefa a ser empregada deve ser norteadada pelo problema em questão e o objetivo almejado.

A tarefa de agrupamento consiste em organizar um conjunto de dados em vários grupos, de modo que, sob alguma definição de “similaridade”, itens semelhantes pertençam ao mesmo grupo e itens diferentes pertençam a grupos distintos (GUHA; MISHRA, 2016).

Ante as potencialidades de aplicação, a análise de grupos tem sido amplamente utilizada em diversos campos. Dentre outros, é possível frisar as áreas de psicologia, das ciências biológicas, da estatística, do reconhecimento de padrões, da recuperação de informações, da aprendizagem de máquina e da mineração de dados (TAN et al., 2006).

Dessa forma, o agrupamento de dados mostrou-se pertinente ao escopo da presente pesquisa, que não persegue a mera classificação, rotulação ou o prenúncio do valor de uma variável. Antes, busca-se segmentar a base em grupos de instâncias semelhantes, no desígnio de prover um ponto de partida para melhor conhecer os dados e gerar uma forma eficiente de explorá-los.

## 1.1 Motivação

No ano de 2018, foram registradas mais de 600 mil CATs e foram gastos mais de R\$13 bilhões em benefícios e indenizações relacionados aos acidentes de trabalho (BRASIL, 2018). Informações como essas podem ser extraídas dos dados brutos relativos aos acidentes de trabalho (CATWEB) e aos benefícios previdenciários (SISBEN), disponíveis no Observatório Digital (OIT, 2017).

A CATWEB é a base de dados composta por todas as Comunicações de Acidentes de Trabalho registradas entre os anos de 2012 e 2017. Essa base possui algumas características peculiares, como: seu grande volume de dados, a ampla diversidade de acidentes de trabalho, a grande quantidade de valores ausentes e a heterogeneidade entre as regiões do país.

Torna-se inviável, portanto, a análise manual dos dados contidos na CATWEB, de modo que a aplicação de técnicas de extração de conhecimento se faz notória. A Mineração de Dados emerge, assim, como mecanismo apto a descobrir, de forma automática ou semiautomática, informações “escondidas” na grande quantidade de dados armazenados, mediante a possibilidade de conferir agilidade ao processo de obtenção de informação (CARDOSO; MACHADO, 2008).

Visando transformar esses dados brutos em informações relevantes, este trabalho busca prover um ponto de partida para a criação de políticas públicas que visem à prevenção de acidentes e doenças ocupacionais, diminuindo o gasto previdenciário e o transtorno para o segurado.

## 1.2 Objetivos e Desafios da Pesquisa

O objetivo geral da pesquisa consiste em reconhecer e caracterizar grupos nos dados relativos aos acidentes de trabalho (CATWEB), a fim de conferir interpretabilidade aos resultados obtidos e extrair informações que subsidiem a tomada de decisão dos gestores públicos.

Para alcançar o objetivo geral, faz-se necessário estabelecer alguns objetivos específicos, tais como:

- ❑ investigar a possível existência de estruturas de grupos nos dados do Ministério Público do Trabalho;
- ❑ implementar uma ferramenta capaz de realizar o pré-processamento dos dados brutos;
- ❑ adaptar o cálculo de medidas de similaridade/dissimilaridade para os diferentes tipos de dados presentes nas bases (categóricos e numéricos);
- ❑ efetuar adaptações na medida de validação Silhueta Simplificada;
- ❑ fornecer grupos que poderão ser melhor investigados por meio de técnicas de visualização de dados;

### 1.3 Hipóteses

Este trabalho pretende ratificar as seguintes hipóteses:

- ❑ **H1:** Usar algoritmos de agrupamento baseados em densidade, como o HDBSCAN\*, produz melhores resultados do que usar algoritmos de agrupamento baseados em probabilidades condicionais, como o CobWeb, para a base em questão.
- ❑ **H2:** Utilizando algoritmos de agrupamento, é possível encontrar grupos de alta relevância, assim avaliados por uma medida de validação, nos dados relativos aos acidentes de trabalho.

### 1.4 Contribuições

As contribuições deste trabalho incluem:

- ❑ A disponibilização de um método para a condução de experimentos envolvendo o agrupamento de dados;
- ❑ Uma ferramenta capaz de realizar o pré-processamento dos dados da base CATWEB;
- ❑ Uma medida de cálculo de distância adaptada para calcular a distância entre duas instâncias da base CATWEB (considerando os seus atributos numéricos e categóricos). Isso possibilitou a execução de algoritmos relacionais, como o HDBSCAN\*, e o cálculo de medidas de validação, que usam medidas de distância entre instâncias, como a Silhueta Simplificada, nos dados da base em questão;
- ❑ A caracterização dos maiores grupos de acidentes de trabalho.

## 1.5 Organização do Trabalho

O presente trabalho está disposto da seguinte forma:

- ❑ **Capítulo 2:** Apresenta uma fundamentação teórica para melhor compreensão desta pesquisa;
- ❑ **Capítulo 3:** Relaciona os trabalhos já desenvolvidos com o trabalho aqui proposto, realçando semelhanças e diferenças entre os mesmos;
- ❑ **Capítulo 4:** Descreve cada etapa do método proposto para o desenvolvimento deste trabalho.
- ❑ **Capítulo 5:** Detalha os experimentos realizados utilizando o algoritmo HDBSCAN\* e apresenta uma análise dos resultados obtidos;
- ❑ **Capítulo 6:** Detalha os experimentos realizados utilizando o algoritmo Cobweb e apresenta uma análise dos resultados obtidos;
- ❑ **Capítulo 7:** Tece as conclusões da pesquisa e indica possíveis trabalhos futuros.

---

## Fundamentação Teórica

Neste capítulo, serão apresentados os conceitos teóricos necessários para melhor compreensão desta pesquisa. As seções 2.1 e 2.2 descrevem, respectivamente, o que é a Comunicação de Acidente de Trabalho (CAT) e como o Ministério Público do Trabalho tem agido de forma a prevenir a ocorrência de acidentes. A Seção 2.3 apresenta uma visão geral do processo de Descoberta de Conhecimento em Bases de Dados, enquanto a Seção 2.4 aprofunda na tarefa de Mineração de Dados que será aplicada ao estudo. As seções 2.5 e 2.6 conceitualizam os algoritmos HDBSCAN\* e COBWEB, empregados nos experimentos deste trabalho. A Seção 2.7 mostra uma forma de comparar hierarquias e a Seção 2.8 elucida formas de avaliar a qualidade do agrupamento obtido. As considerações finais serão discutidas na Seção 2.9.

### 2.1 Comunicação de Acidente de Trabalho - CAT

O Acidente de Trabalho pode ser um acidente que aconteceu durante o exercício da atividade laboral, doença ocupacional decorrente do exercício do trabalho (ou a ele relacionada) ou um acidente ocorrido no trajeto até o trabalho (BRASIL, 2019).

A CAT é um documento com a finalidade de registrar tanto um acidente de trabalho ou de trajeto, como uma doença ocupacional. Deve ser feito pelo empregador, mesmo que não haja afastamento do acidentado (INSS, 2018).

Em uma CAT, constam as informações sobre o empregado (nome, sexo, idade, cargo), o empregador e o acidente (local, data, lesão, tipo de acidente e agente causador).

Conforme disposto pela Comissão Interna de Prevenção de Acidentes (CIPA, 2011), uma Comunicação de Acidente de Trabalho pode ser diferenciada em três tipos, são eles:

- ❑ CAT inicial: primeira comunicação de um acidente típico de trabalho ou de trajeto, ou ainda, de uma doença ocupacional;
- ❑ CAT de reabertura: reinício de tratamento ou agravamento de uma lesão ou doença que já foram comunicados anteriormente ao INSS;

- CAT de comunicação de óbito: emitido para falecimento decorrente de acidente ou doença profissional ou do trabalho, após a emissão da CAT inicial.

Com a finalidade de facilitar o registro da CAT, o INSS tornou disponível um aplicativo que permite executar todo o processo *online*, ou ainda, é possível imprimir o formulário em branco, preenchê-lo de forma manual e registrá-lo em uma agência do INSS. No entanto, para que isso seja possível, é necessário que todos os campos obrigatórios sejam devidamente preenchidos (INSS, 2018).

Para o atendimento em qualquer agência do INSS, é indispensável a apresentação de um documento de identificação com foto e o número do CPF. Ademais, a CAT deverá ser emitida em quatro vias, sendo: 1ª via do INSS, 2ª via do segurado ou dependente, 3ª via do sindicato da classe do trabalhador e a 4ª via da empresa (BARTOLOMEU, 2002).

O registro da CAT é de fundamental importância, não somente para assegurar ao acidentado os direitos que lhe são devidos, como também para fomentar bases de dados que poderão ser utilizadas para pesquisas (SOUSA et al., 2006; BARTOLOMEU, 2002; SILVA, 2018).

A base de dados adotada ao longo deste trabalho é composta por Comunicações de Acidente de Trabalho registradas entre os anos de 2012 a 2017. Ressaltando que cada registro armazena os dados de uma CAT de forma adequadamente anônima.

## 2.2 Medidas do MPT para prevenir acidentes

No Brasil, a cada quatro horas e meia, morre um trabalhador vítima de acidente de trabalho (RODRIGUES, 2018). Com o propósito de prevenir a ocorrência de acidentes e doenças ocupacionais, o Ministério Público do Trabalho (MPT) tem promovido diversas ações, como campanhas, seminários, palestras e cartilhas educativas.

Em abril de 2018, o MPT lançou a campanha denominada “Abril Verde”, cuja ideia era a implantação de uma nova cultura para a prevenção de acidentes de trabalho e doenças ocupacionais, de forma a garantir segurança, tranquilidade e melhores condições trabalhistas para todos os brasileiros (LOBO, 2018). Para divulgação da campanha, diversos prédios, monumentos e espaços foram iluminados de verde. Essa prática já é tradicional em outras campanhas, como o Outubro Rosa, voltada para a prevenção do câncer de mama, e o Novembro Azul, criada para alertar sobre o câncer de próstata (FRANCO, 2018).

Ainda em abril, dentro do contexto da campanha supracitada, ocorreu um evento em Campo Grande - MS, para discutir o adoecimento ocupacional e a gestão de riscos para o trabalho em altura. Temas como o “Impacto da Gestão na Saúde dos Trabalhadores” e “Gestão de Riscos para Trabalho em Altura” foram abordados em palestras, a fim de sensibilizar a sociedade e conscientizá-la quanto à importância da prevenção de acidentes e doenças do trabalho (MPT-MS, 2018).

Outrossim, foram lançadas cartilhas com orientações para os trabalhadores, como o “Guia Básico de Prevenção de Acidentes em Espaços Confinados: poços e cisternas”, ou ainda, a cartilha “Obra Legal” - no contexto da área de construção civil (MPT-PB, 2018).

Sob a perspectiva do MPT, as empresas devem se atentar às previsões legais sobre saúde e segurança do trabalho, fornecer os equipamentos de proteção individual (EPI's) necessários e manter um ambiente laboral seguro. É importante ressaltar, também, a necessidade de ampliar a fiscalização por parte do Ministério Público do Trabalho (FRANCO, 2018).

Por fim, uma recente iniciativa do MPT visando à prevenção de acidentes é o Observatório Digital de Saúde e Segurança do Trabalho, que será mais bem descrito na Subseção 2.2.1.

### 2.2.1 Observatório Digital de Saúde e Segurança do Trabalho

Em 2017, o Ministério Público do Trabalho, em parceria com a Organização Internacional do Trabalho (OIT)<sup>1</sup> iniciaram um fórum multidisciplinar para estimular a adoção de práticas inteligentes de gestão do conhecimento orientada para impactos sociais (*Smartlab* de Trabalho Decente). Nesse contexto, criaram o Observatório Digital de Saúde e Segurança do Trabalho<sup>2</sup>.

O observatório é uma ferramenta *online*, com o objetivo de melhorar o acesso à informação e potencializar projetos, programas e políticas públicas de Trabalho Decente e fortalecer a atuação do Ministério Público do Trabalho e de outras instituições parceiras, contribuindo com a construção de medidas de avaliação da eficiência e da efetividade das ações executadas (ERPLAN, 2018).

Na plataforma, consta a informação sobre os bancos de dados que sustentam o Observatório Digital, são eles:

- ❑ Relação Anual de Informações Sociais (RAIS), Ministério do Trabalho.
- ❑ Cadastro Geral de Empregados e Desempregados (CAGED), Ministério do Trabalho.
- ❑ Sistema Único de Informações de Benefícios da Previdência Social (SISBEN), Ministério da Fazenda.
- ❑ Pesquisa Nacional por Amostra de Domicílios (PNAD), IBGE.
- ❑ Censo, IBGE.
- ❑ Sistema de Indicadores Municipais de Trabalho Decente, OIT.

<sup>1</sup> <https://www.ilo.org/brasil/comece-a-oit/lang-pt/index.htm>

<sup>2</sup> <https://smartlabbr.org/sst>

#### □ IPEADATA, IPEA.

Entre os dados apresentados, ressaltam-se indicadores de frequência de acidentes de trabalho, quantidade de CAT's registrados, gastos previdenciários, dias de trabalho perdidos, indicativo de óbitos, localização dos acidentes e afastamentos, ramos de atividade econômica envolvidos, perfis das vítimas e Classificação Internacional de Doenças.

Embora o Observatório Digital seja uma iniciativa interessante, apenas a visualização e a sumarização dos dados não são suficientes. Segundo Brito (2019), o processo de exploração e a análise de dados disponíveis no site são muito limitados e não são adequados para revelar padrões complexos. Desta forma, acredita-se que técnicas capazes de identificar relações ou grupos nos dados automaticamente, poderiam ajudar a responder a questões importantes, tais como: o perfil dos trabalhadores que mais se acidentam, os setores que oferecem mais risco à saúde, se existem relações entre acidentes ocorridos em diferentes regiões geográficas, dentre outras.

## 2.3 Descoberta de Conhecimento em Bases de Dados

O termo *Knowledge Discovery in Databases* (KDD), ou Descoberta de Conhecimento em Bases de Dados, refere-se a todo o processo de extração de conhecimento a partir dos dados. Enquanto a Mineração de Dados refere-se apenas à aplicação de algoritmos para buscar padrões, o KDD engloba, ainda, a interpretação e avaliação dos padrões encontrados, ponderando sobre o que realmente constitui conhecimento útil (RAMOS; LOBO, 2003).

O KDD é um processo não-trivial, para identificação de padrões compreensíveis, novos, válidos e potencialmente úteis a partir de grandes bases de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Porquanto padrões compreensíveis são representações do conhecimento que podem ser interpretadas por seres humanos (GOLDSCHMIDT; PASSOS, 2015). A Figura 1 ilustra as principais etapas envolvidas no processo de Descoberta de Conhecimento em Bases de Dados, que serão mais bem detalhadas nas próximas seções.

Sob a perspectiva da realização do processo, o KDD consiste em uma sequência de complexas interações, entre um usuário e uma coleção de dados, normalmente auxiliada por um conjunto heterogêneo de ferramentas computacionais (BRACHMAN; ANAND, 1996).



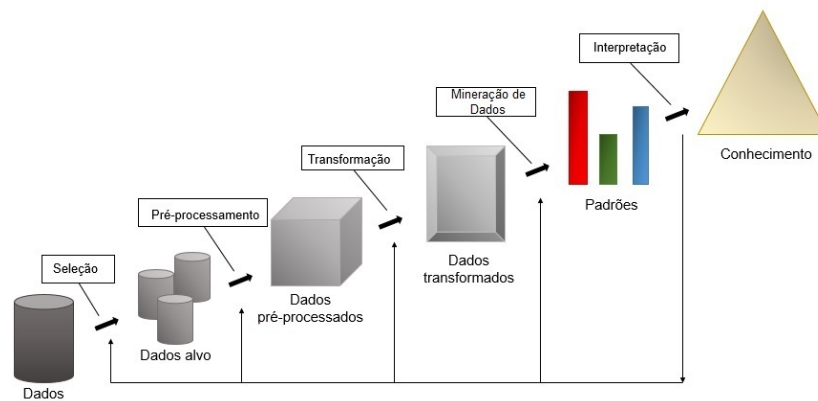


Figura 1 – Etapas do processo de KDD.

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

### 2.3.1 Seleção dos dados

A etapa de seleção de dados consiste em definir os dados (subconjuntos ou amostras) que deverão ser efetivamente considerados durante o KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Existem duas abordagens distintas para a seleção de dados, são elas: a seleção de atributos e a seleção de registros (GOLDSCHMIDT; PASSOS, 2015).

Na seleção de atributos, características redundantes ou irrelevantes ao KDD são desconsideradas, reduzindo a dimensionalidade do conjunto de dados (TAN et al., 2006). O raio e o diâmetro de um círculo são exemplos de informações redundantes, assim como o número de matrícula e o sexo do aluno são atributos irrelevantes, quando se deseja avaliar o desempenho médio de uma turma em determinada disciplina.

Já a seleção de objetos ou instâncias é empregada quando não é possível trabalhar com todo o conjunto de dados ou quando se deseja otimizar a análise de mineração de dados quanto ao tempo ou custo. Nesse contexto, podem ser empregadas técnicas de agregação (combinação de dois ou mais instâncias em uma única) ou amostragem (subconjunto de instâncias) (TAN et al., 2006).

### 2.3.2 Pré-processamento

O pré-processamento compreende a aplicação de técnicas para conferir qualidade (completude, veracidade e integridade) aos dados. Nesta etapa, são efetuados processos para a eliminação de ruídos, tratamento de valores ausentes e de informações inconsistentes (GOLDSCHMIDT; PASSOS, 2015).

As tarefas realizadas durante o pré-processamento podem ser fortemente dependentes de conhecimento de domínio, quando necessitam de um conhecimento específico para serem aplicadas, ou fracamente dependentes de conhecimento de domínio, quando as informações necessárias para tratar o problema de pré-processamento são extraídas dos

próprios dados (BATISTA, 2003). A verificação de integridade é um tipo de tarefa que requer conhecimento de domínio, enquanto o tratamento de valores ausentes pode ser feito de forma mais automatizada.

### 2.3.3 Transformação

Na etapa de transformação, os dados são modificados de modo a assumir o formato de entrada apropriado para os algoritmos de mineração que serão utilizados (GOLDSCHMIDT; PASSOS, 2015). As técnicas empregadas deverão ser escolhidas de acordo com o objetivo pretendido.

Uma das motivações para se usar técnicas de transformação de dados é o grande número de valores para um atributo no conjunto de dados. Além disso, os valores podem estar em formato inadequado para a aplicação direta do algoritmo de mineração de dados. São exemplos de transformação: agregação, normalização, generalização, codificação e criação de novos atributos (BRAGA, 2005).

### 2.3.4 Mineração de Dados

A Mineração de Dados (MD), do inglês *Data Mining*, é a principal etapa dentro do processo de KDD, e consiste na aplicação de algoritmos de análise e descoberta de dados, para a extração de padrões (FAYYAD et al., 1996).

Sob uma outra perspectiva, voltada para a estatística, Mineração de Dados pode ser definida como uma análise de grandes conjuntos de dados, com a finalidade de encontrar relacionamentos inesperados e resumir os dados de forma que sejam úteis e compreensíveis (HAND, 2007).

As tarefas de Mineração de Dados podem ser divididas em duas categorias:

- ❑ **Previsão:** tem o objetivo de prever o valor de uma variável (alvo), baseado nos valores de outras variáveis (explicativas).
- ❑ **Descrição:** centra-se na busca por padrões (grupos, tendências, correlações) que resumam os relacionamentos subjacentes nos dados. Agrupamento, sumarização e associação, são exemplos de tarefas descritivas.

Algumas das tarefas mais comuns da Mineração de Dados são: classificação, regressão, associação, sumarização e agrupamento. A seguir, serão abordados os conceitos por trás de cada uma.

Na tarefa de classificação, o objetivo é identificar a qual classe um determinado objeto pertence. Para isso, um modelo é treinado com um conjunto de registros rotulados, com a finalidade de aprender como classificar um novo registro (aprendizado supervisionado)(BRAGA, 2005).

A regressão é semelhante à classificação, porém é utilizada para valores numéricos, enquanto a classificação prevê valores categóricos. O objetivo desta tarefa é buscar uma função que faça mapeamento dos dados em variáveis de valor real (FAYYAD et al., 1996). Tanto a classificação quanto a regressão são tarefas do tipo preditivas.

A associação é usada para encontrar características (atributos) altamente associadas dentro dos dados. Sendo que, tais características podem ser representadas na forma de regras de implicação (**se** atributo x, **então** atributo y) (TAN et al., 2006).

No contexto da Mineração de dados, a sumarização tem o objetivo de fornecer descrições compactas dos subconjuntos de dados, sendo comumente utilizada para a geração automática de relatórios (RAMOS; LOBO, 2003).

Agrupamento, do inglês *Clustering*, é a tarefa que busca identificar grupos de instâncias similares. De tal forma que, as instâncias dentro um mesmo grupo devem ser mais semelhantes entre si e mais dissemelhantes às instâncias de outros grupos (TAN et al., 2006).

As tarefas de associação, sumarização e agrupamento, são do tipo descritivas. A Subseção 2.4 dará maior ênfase na tarefa de agrupamento, pois será a utilizada na fase experimental deste estudo.

### 2.3.5 Interpretação

A última etapa do processo de KDD envolve a análise, visualização e interpretação dos padrões encontrados na etapa de Mineração de Dados. Embora os algoritmos de Mineração de Dados automatizem boa parte do processo de descoberta de conhecimento, os resultados, ainda requerem a análise humana (BRAGA, 2005).

Todos os envolvidos no projeto devem participar da etapa de interpretação, tanto os especialistas em KDD, quanto os especialistas de domínio da aplicação, pois deverá ser realizada uma avaliação dos resultados obtidos e, caso seja devido, a definição de novas táticas para investigação dos dados (BOENTE; ROSA, 2007).

Finalmente, em posse do conhecimento consolidado, existem três tipos diferentes de ações (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996): i) usar o conhecimento diretamente para a tomada de decisões; ii) incorporação do conhecimento em outro sistema, para ações futuras; iii) simplesmente documentar o conhecimento e relatar aos interessados.

## 2.4 Agrupamento

Agrupamento pode ser definido como um processo que, usando apenas as informações intrínsecas de uma base de dados, descobre um conjunto de partições que representa sua estrutura inerente (ALONSO, 2013).

Existem diversos tipos de algoritmos de agrupamento na literatura, sendo que os mais comuns são os hierárquicos, particionais e os baseados em densidade (TAN et al., 2006). O algoritmo deve ser escolhido levando-se em consideração os tipos dos dados e o objetivo pretendido.

As subseções 2.4.2, 2.4.3, 2.4.4 explicitam uma visão geral dos algoritmos de agrupamento baseados em métodos particionais, hierárquicos e de densidade, respectivamente. No entanto, antes de apresentar os principais tipos de algoritmos, considerou-se importante discutir a questão sobre similaridade entre os dados. Para isso, na Subseção 2.4.1 foram apresentadas algumas medidas de proximidade.

### 2.4.1 Medidas de proximidade

Um algoritmo de agrupamento tem como finalidade encontrar um agrupamento no qual as instâncias dentro de um grupo sejam mais semelhantes entre si do que em relação às instâncias de outros grupos (NALDI, 2011). Antes de inferir quão semelhantes ou distintas são as instâncias, é preciso determinar como as medidas de semelhança ou diferença serão calculadas (AGGARWAL, 2014).

No contexto deste trabalho, optou-se por adotar o critério de dissimilaridade, sendo que a forma definida para mensurar a dissimilaridade entre os dados foi a distância. Segundo Frei (2006), quanto maior o valor da dissimilaridade, menor a semelhança entre as instâncias.

Para definir o quão dissimilares duas instâncias são, qualquer medida de distância pode ser empregada, sendo necessário considerar a natureza dos dados (discreta, contínua ou binária), as escalas de medida (nominal, ordinal, intervalar ou racional) e o conhecimento sobre do assunto em questão (KASZNAR; GONÇALVES; BENTO, 2009).

Como observado por Santos (2018) em seu trabalho, a distância entre duas instâncias deve seguir quatro princípios fundamentais:

- **Não-negatividade:** seja  $d(a, b)$  a distância entre as instâncias  $a$  e  $b$ , então  $d(a, b) \geq 0$ ;
- **Reflexividade:**  $d(a, b) = 0$  se e somente se,  $a = b$ ;
- **Simetria:**  $d(a, b) = d(b, a)$ ;
- **Desigualdade triangular:**  $d(a, b) + d(b, c) \geq d(a, c)$

A Distância de Minkowski, definida pela equação 1, é uma métrica para o espaço Euclidiano que serve como generalização para outras distâncias, tais como: City Block, Euclidiana e Chebyshev (JAIN; MURTY; FLYNN, 1999).

$$d(x, y) = \left( \sum_{i=1}^n |x_i - y_i|^k \right)^{\frac{1}{k}} \quad (1)$$

onde:  $n$  é o tamanho do vetor de atributos (ou ainda, o número de atributos) e  $k$  vai definir o tipo de distância que se deseja calcular. Sendo que, para  $k = 1$  calcula-se a Distância City Block ou Manhattan,  $k = 2$  obtem-se a Distância Euclidiana, e, por fim, adotando  $k = \infty$  será possível mensurar a Distância Chebyshev ou Chessboard.

O caso mais recorrentemente utilizado dessa distância é a Euclidiana, quando  $k = 2$  e corresponde à distância geométrica no espaço multidimensional.

#### 2.4.1.1 Distância em dados categóricos

Os atributos categóricos podem ser subdivididos em binários, nominais e ordinais. Os atributos binários possuem apenas dois valores possíveis, como: {verdadeiro ou falso} e {0 ou 1}. Já os atributos nominais são caracterizados por símbolos distintos, para definir a cor dos olhos, por exemplo, tem-se os seguintes valores: {azul, verde e castanho}. Por fim, os atributos ordinais são representados por valores que indicam uma ordem entre eles, como: {pequeno, médio e grande} (TAN et al., 2006).

Ao se deparar com atributos categóricos, não basta aplicar a fórmula da distância apresentada na Equação 1. Para tratar desse caso em particular, existem duas abordagens distintas: converter o atributo categórico ou buscar alternativas capazes de lidar com esse tipo de dado.

Uma forma de converter o atributo categórico é o método de codificação inteira-binária, onde deve-se associar a cada valor de um atributo, um número inteiro no intervalo de  $[0, m - 1]$ , respeitando a ordem no caso do valor ser ordinal. Em seguida, cada um dos  $m$  inteiros é convertido para binário. A Tabela 1 ilustra um exemplo dessa codificação.

Atributo categórico	Número inteiro	$x_1$	$x_2$	$x_3$
Horrível	0	0	0	0
Ruim	1	0	0	1
Médio	2	0	1	0
Bom	3	0	1	1
Ótimo	4	1	0	0

Tabela 1 – Codificação inteira-binária  
Fonte: adaptado de Tan et al. (2006)

Segundo (TAN et al., 2006), um problemas desse método é o fato de criar relacionamento não pretendido entre os atributos convertidos, por exemplo, os atributos  $x_2$  e  $x_3$  da Tabela 2, são correlacionados para o valor “Bom”, o que não necessariamente corresponde à realidade. Um outro problem observado é que a distância entre “Horrível” e “Ruim” e a distância entre “Horrível” e “Ótimo” serão iguais, perdendo a ordem entre esses valores.

Existe, ainda, uma outra forma de converter o atributo categórico para numérico, trata-se da conversão 1-de-n. Nesse método é criado um atributo para cada valor possível. A Tabela 2 mostra um exemplo de conversão 1-de-n, onde  $x_1$  representa o valor “Horrível”,

$x_2$  representa o valor “Ruim”,  $x_3$  representa o valor “Médio”,  $x_4$  representa o valor “Bom” e  $x_5$  representa o valor “Ótimo”.

Atributo categórico	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
Horrível	1	0	0	0	0
Ruim	0	1	0	0	0
Médio	0	0	1	0	0
Bom	0	0	0	1	0
Ótimo	0	0	0	0	1

Tabela 2 – Codificação 1-de-n  
Fonte: adaptado de Tan et al. (2006)

A grande desvantagem desse método é que, se um atributo possuir muitos valores distintos, serão criados muitos novos atributos. Isso pode implicar o problema chamado de maldição da dimensionalidade, problema causado pelo aumento exponencial no volume associado com a adição de dimensões extras a um espaço matemático (BELLMAN, 2015).

Finalmente, existe a opção adotada neste trabalho, que é a não-conversão do atributo categórico. Desta forma, assume-se que: se dois atributos categóricos nominais forem exatamente iguais, a distância entre eles é 0, caso contrário, a distância é 1.

## 2.4.2 Métodos particionais

Os algoritmos que empregam o método particional simplesmente realizam uma divisão do conjunto de dados em  $k$  grupos não interseccionados, sendo que o número de grupos deve ser previamente informado pelo usuário (MENDES, 2017).

Os algoritmos de agrupamento buscam minimizar a distância intercluster e maximizar a distância intracluster (TAN et al., 2006), ou seja, exemplos pertencentes ao mesmo grupo devem ser mais similares entre si e menos similares aos exemplos pertencentes a grupos distintos.

O *k-means* (MACQUEEN et al., 1967) é um dos algoritmos particionais mais conhecidos. A popularidade deste algoritmo se deve a sua simplicidade, escalabilidade e sucesso empírico (WU et al., 2008). A ideia geral do algoritmo *k-means* é:

- ❑ Escolher  $k$  centroides iniciais, onde  $k$  é o parâmetro especificado pelo usuário, que indica a quantidade desejada de grupos;
- ❑ Atribuir cada objeto do conjunto de dados ao centroide mais próximo, baseando-se em alguma medida de similaridade (Euclidiana, por exemplo). A coleção de instâncias atribuídas a um centroide, constitui um grupo;
- ❑ O centroide de cada grupo deve ser atualizado, considerando todas as instâncias atribuídas aos grupos;

- Os passos 2 e 3 devem ser repetidos até que nenhuma mudança ocorra, ou seja, os *clusters* se estabilizem.

As seguintes desvantagens do *k-means* são enfatizadas: o algoritmo é sensível a ruídos e *outliers* e encontra apenas grupos de formato hipersférico. Além disso, é preciso fornecer o número de grupos como parâmetro e o algoritmo trabalha apenas com dados numéricos (BORGES, 2010).

### 2.4.3 Métodos hierárquicos

No agrupamento do tipo hierárquico, os grupos representam a junção de seus subgrupos, seguindo a estrutura de uma árvore ou dendrograma. Cada nó é um *cluster* e a raiz é o grupo contendo todos os elementos. As folhas, geralmente, representam as instâncias de dados individuais, mas isso não é uma regra (FREI, 2006).

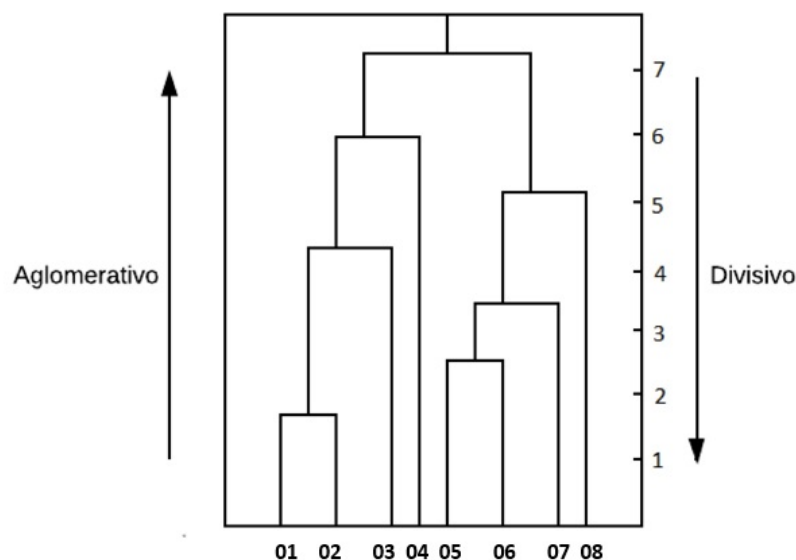


Figura 2 – Dendrograma obtido pelo método de agrupamento hierárquico.

Fonte: adaptado de Silva (2018)

Os métodos hierárquicos podem usar duas abordagens distintas: aglomerativa ou divisiva, conforme Figura 2. As estratégias divisivas começam com o conjunto de dados inteiro e, em cada iteração, é determinada uma maneira de dividir os dados em duas partições. Este processo é repetido recursivamente até que exemplos individuais sejam alcançados ou algum critério de parada seja satisfeito. Por outro lado, as estratégias aglomerativas reúnem iterativamente o par de partições mais relacionado, de acordo com uma medida de similaridade, até que haja apenas uma partição (ALONSO, 2013).

Em geral, os algoritmos do tipo hierárquicos não precisam armazenar os registros de dados, apenas as distâncias (similaridade/dissimilaridade) entre eles. Ademais, são algoritmos que não lidam bem com ruídos e *outliers* e são sensíveis à ordem de entrada dos

dados. Em contrapartida, não requerem conhecimento prévio sobre o número de grupos e seu resultado corresponde a uma taxonomia dos dados, podendo fornecer informações importantes sobre o conjunto de dados (ARRUDA, 2011).

O *Single-linkage* (FLOREK K., 1951) é um exemplo de algoritmo de agrupamento hierárquico aglomerativo baseado no conceito de vizinhos mais próximos, no qual se considera que a similaridade entre as duas instâncias mais similares de dois determinados grupos corresponde à similaridade entre tais grupos (LINDEN, 2009), conforme a Equação 2.

$$d(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y) \quad (2)$$

Desta forma  $d(C_1, C_2)$  representa a distância entre os grupos  $C_1$  e  $C_2$ , dada pela distância mínima entre as duas instâncias mais próximas,  $x$  e  $y$ , pertencentes aos grupos  $C_1$  e  $C_2$ , respectivamente.

Segundo Santos (2018), o *Single-linkage* é capaz de encontrar grupos com formas arbitrárias e complexas nos dados. No entanto, trata-se de um algoritmo sensível à presença de ruídos.

#### 2.4.4 Métodos baseados em densidade

Algoritmos de agrupamento baseados em densidade têm como objetivo identificar áreas de alta densidade, que estejam rodeadas por regiões de baixa densidade (SEMAAN, 2013). Esses métodos apresentam como principais vantagens: a possibilidade de descobrir grupos de formatos arbitrários, a capacidade de tratar ruídos e o fato de que o número de grupos não precisa ser conhecido *a priori* (ALONSO, 2013). Exemplos de algoritmos baseados em densidade: DBSCAN (ESTER et al., 1996), OPTICS (ANKERST et al., 1999), DENCLUE (HINNEBURG; KEIM, 2003).

Aqui será dado destaque ao algoritmo DBSCAN, pois este é de fundamental importância para a compreensão de um dos algoritmos adotados nesta pesquisa, o HDBSCAN\*. Ressalta-se que serão apresentadas as definições formais do DBSCAN tal como descritas em Ester et al. (1996).

Seja  $X$  um conjunto de dados multidimensionais,  $x$  e  $y$  objetos pertencentes ao conjunto  $X$ , têm-se que:

**Definição 1.  $\varepsilon$  - vizinhança de um objeto:** A  $\varepsilon_{vizinhança}$ , denotada por  $N_\varepsilon(x)$ , é definida por  $N_\varepsilon(x) = \{y \in X | d(x, y) \leq \varepsilon\}$ .

Os objetos podem ser classificados em três tipos diferentes: objeto *core*, objeto *border* ou ruído. Objetos *core* precisam ter uma quantidade mínima de objetos ( $m_{pts}$ ) em suas  $\varepsilon_{vizinhanças}$ . Objetos *border* não contêm pelo menos uma quantidade mínima de objetos em suas  $\varepsilon_{vizinhanças}$ , mas pertencem a  $\varepsilon_{vizinhanças}$  de pelo menos um objeto *core*. Por fim, os ruídos não contêm  $m_{pts}$  objetos em suas  $\varepsilon_{vizinhanças}$  e não são vizinhos de nenhum objeto *core*.



**Definição 2. Diretamente acessível:** Diz-se que um objeto  $x$  é diretamente acessível por um objeto  $y$ , com relação a  $\varepsilon$  e  $m_{pts}$ , se:

1.  $x \in N_\varepsilon(y)$
2.  $|N_\varepsilon(y)| \geq m_{pts}$

Segundo Ester et al. (1996), a definição de diretamente acessível é simétrica para pares de objetos *core*.

**Definição 3. Densamente acessível:** Um objeto  $x$  é densamente acessível por um objeto  $y$ , com relação a  $\varepsilon$  e  $m_{pts}$ , se existe uma sequência de objetos  $x_1, \dots, x_n, x_1 = y, x_n = x$ , em que  $x_{i+1}$  é diretamente acessível por  $x_i$ .

**Definição 4. Densamente conectado:** Um objeto  $x$  é densamente conectado a um objeto  $y$ , com relação a  $\varepsilon$  e  $m_{pts}$ , se há um objeto  $z$  tal que  $x$  e  $y$ , são densamente acessíveis por  $z$ .

**Definição 5. Grupo:** Um grupo  $C$ , com relação a  $\varepsilon$  e  $m_{pts}$ , é um subconjunto de  $X$  não vazio que satisfaz as seguintes condições:

1.  $\forall x, y$ : se  $x \in C$  e  $y$  é densamente acessível por  $x$ , com relação a  $\varepsilon$  e  $m_{pts}$ , então  $y \in C$  (Maximalidade).
2.  $\forall x, y \in C$  :  $x$  é densamente conectado a  $y$ , com relação a  $\varepsilon$  e  $m_{pts}$  (Conectividade).

**Definição 6. Ruído:** É um subconjunto de instâncias em um conjunto de dados  $X$  que não pertencem a nenhum grupo, com relação a  $\varepsilon$  e  $m_{pts}$ .

Os lemas a seguir são importantes para validar a correção do algoritmo de agrupamento proposto por Ester et al. (1996). Dados os parâmetros  $\varepsilon$  e  $m_{pts}$ , é possível descobrir um grupo em uma abordagem em duas etapas. Primeiro, escolher uma instância aleatoriamente no conjunto de dados  $X$ , que satisfaça à condição do objeto *core*, como uma semente inicial. Segundo, recuperar todos os objetos que são densamente acessíveis por essa semente, obtendo o grupo que a contém.

**Lema 1:** Seja  $x$  um objeto pertencente a  $X$  e  $N_\varepsilon(x) \geq m_{pts}$ . Então, o conjunto  $O = \{o | o \in X \text{ e } o \text{ é densamente acessível por } x, \text{ em relação a } \varepsilon \text{ e a } m_{pts}\}$  é um grupo.

**Lema 2:** Seja  $C$  um grupo, com relação a  $\varepsilon$  e a  $m_{pts}$ , e seja  $x$  um objeto pertencente a  $C$  com  $|N_\varepsilon(x)| \geq m_{pts}$ . Então,  $C$  é igual ao conjunto  $O = \{o | o \text{ é densamente acessível por } x, \text{ com relação a } \varepsilon \text{ e a } m_{pts}\}$ .

Algumas desvantagens desse algoritmo são: a sensibilidade em relação aos parâmetros de entrada e a dificuldade em trabalhar com instâncias com densidades muito variadas (TAN et al., 2006).

## 2.5 HDBSCAN\*

Um algoritmo recente, considerado estado da arte em agrupamento hierárquico, é o HDBSCAN\*. Porém, antes de apresentá-lo, faz-se necessário ressaltar algumas definições do DBSCAN\*.

Destaca-se que os conceitos e definições aqui apresentados, tanto do algoritmo DBSCAN\* quanto do HDBSCAN\*, são fundamentados nos trabalhos de Campello, Moulavi e Sander (2013), Campello et al. (2015) e Santos (2018).

### 2.5.1 Definições do DBSCAN\*

Segundo Campello et al. (2015), o diferencial do DBSCAN\* em relação ao DBSCAN proposto por Ester et al. (1996), é que os grupos não são definidos apenas com base nos objetos *core*.

Seja um conjunto de dados contendo  $n$  objetos,  $X = \{x_1, x_2, \dots, x_n\}$ , e  $D$  uma matriz de distância de tamanho  $n * n$ , com a distância entre os pares de objetos  $d(x_p, x_q)$ , em que  $x_p$  e  $x_q \in X$ . Ressalta-se que a matriz  $D$  não é necessária, desde que as distâncias dos objetos em  $X$  possam ser calculadas sob demanda.

**Definição 1. Objeto *core*:** um objeto  $x_p$  é um objeto *core*, com relação aos parâmetros  $\epsilon$  e  $m_{pts}$ , se na sua vizinhança  $\epsilon_{vizinhança}$  contém ao menos  $m_{pts}$  objetos. Os objetos que não são *core* são chamados de ruídos.

**Definição 2. Objetos  $\epsilon_{acessíveis}$ :** dois objetos  $x_p$  e  $x_q$  são  $\epsilon_{acessíveis}$ , em relação aos parâmetros  $\epsilon$  e  $m_{pts}$ , se  $x_p \in N_\epsilon(x_q)$  e  $x_q \in N_\epsilon(x_p)$ , onde  $N_\epsilon(x_q)$  é a vizinhança de  $x_q$  e  $N_\epsilon(x_p)$  é a vizinhança de  $x_p$ .

**Definição 3. Objetos densamente conectados:** dois objetos são densamente conectados se eles são direta ou transitivamente acessíveis, em relação aos parâmetros  $\epsilon$  e  $m_{pts}$ .

**Definição 4. Grupo:** Um grupo  $C$ , em relação aos parâmetros  $\epsilon$  e  $m_{pts}$ , é um subconjunto máximo não vazio de  $X$ , em que todo par de objetos em  $C$  é densamente conectado.

O algoritmo DBSCAN\* considera que todos os objetos de um conjunto de dados  $X$  são vértices de um grafo, tal que cada par de vértices é adjacente, se, e somente se, os objetos envolvidos são  $\epsilon_{acessíveis}$  em relação aos parâmetros  $\epsilon$  e  $m_{pts}$ .

### 2.5.2 Definições HDBSCAN\*

**Definição 5. Core distance:** A Core distance  $d_{core}(x_p)$  de um objeto  $x_p \in X$ , com relação a  $m_{pts}$ , é um vetor de distância de  $x_p$  aos seus  $m_{pts}$  vizinhos mais próximos, incluindo  $x_p$ .

Um objeto  $x_p \in X$  é chamado de objeto  $\epsilon$ -core para todo valor de  $\epsilon$  que é maior ou igual à distância *core* de  $x_p$  em relação a  $m_{pts}$ , isto é, se  $d_{core}(x_p) \leq \epsilon$ .

**Definição 6. Distância de Acessibilidade Mútua (*Reachability*):** A distância de acessibilidade mútua  $d_{mreach}$  indica o raio mínimo  $\epsilon$  entre os objetos  $x_p$  e  $x_q$ , tal que ambos sejam  $\epsilon$ -acessíveis. Essa distância é definida pela fórmula:  $d_{mreach} = \max[d_{core}(x_p), d_{core}(x_q), d(x_p, x_q)]$ .

**Definição 7. Grafo de Distância de Acessibilidade Mútua:** É um grafo completo,  $G_{mpts}$ , no qual os objetos de  $X$  são vértices e o peso de cada aresta é a distância de acessibilidade mútua, em relação ao seu  $m_{pts}$ , entre o respectivo par de objetos.

Seja  $G_{mpts, \varepsilon} \subseteq G_{mpts}$  um grafo obtido removendo todas as arestas de  $G_{mpts}$  com pesos maiores que algum valor de raio  $\varepsilon$ . Considerando-se as definições 4 e 8, em relação aos parâmetros  $m_{pts}$  e  $\varepsilon$ , grupos são definidos como os componentes conectados dos objetos *core* em  $G_{mpts, \varepsilon}$ , e os objetos restantes são classificados como ruídos (CAMPELLO; MOULAVI; SANDER, 2013; CAMPELLO et al., 2015).

### 2.5.3 O Algoritmo HDBSCAN\*

O HDBSCAN\*, ou DBSCAN\* hierárquico, configura um algoritmo hierárquico divisivo (*top-down*) baseado na Árvore Geradora Mínima ou *Minimum Spanning Tree (MST)* obtida do Grafo de Distância de Acessibilidade Mútua. Nesse contexto, uma Árvore Geradora Mínima corresponde a um subgrafo acíclico, composto por um conjunto de vértices conectados por arestas ponderadas, na qual a soma total dos pesos dessas arestas seja mínima (GROSS; YELLEN, 2005).

A proposta deste algoritmo é, a partir da MST gerada, remover as arestas de maior peso (distância de acessibilidade mútua) em ordem decrescente, construindo cada nível da hierarquia, composto por objetos conectados (grupos) e objetos isolados (ruídos) (CAMPELLO; MOULAVI; SANDER, 2013; CAMPELLO et al., 2015). O algoritmo 1 ilustra o pseudo-código do HDBSCAN\* que tem como entradas um valor para  $m_{pts}$  e o conjunto de dados  $X$ . Ele produz uma árvore de grupos que contém todas as partições obtidas pelo DBSCAN\* de uma maneira hierárquica e aninhada (CAMPELLO; MOULAVI; SANDER, 2013).

Os autores desse algoritmo observaram que os grupos encontrados podem sofrer três evoluções distintas ao longo da hierarquia, conforme o  $\varepsilon$  é decrementado. São elas:

1. O grupo pode diminuir de tamanho, ou seja, o número de objetos é reduzido, mas ainda é caracterizado como um grupo.
2. O grupo pode se dividir em subgrupos menores.
3. O grupo pode desaparecer à medida que a hierarquia é construída.

**Algoritmo 1** HDBSCAN\***begin**

- 1: Calcula a *Core distance*, em relação a  $m_{pts}$  para todos os objetos de  $X$ ;
- 2: Calcula a MST, a partir do Grafo de Distância de Acessibilidade Mútua ( $G_{mpts}$ );
- 3: Estende a MST para obter MST-estendida, adicionando para cada vértice (objeto) um “*self-loop*” com a distância *core* do objeto correspondente;
- 4: Ordena MST-estendida em ordem decrescente de peso e extrai a hierarquia HDBSCAN\* como um dendrograma do MST-estendida;
  - 4.1: Para a raiz da árvore, atribua a todos os objetos o mesmo rótulo (grupo único);
  - 4.2: Remover iterativamente todas as arestas da MSTestendida em ordem decrescente de pesos;
    - 4.2.1: Antes de cada remoção, defina o valor da escala de dendrograma do nível hierárquico corrente, como o peso das arestas a serem removidas.
    - 4.2.2: Após a remoção, atribuem-se rótulos para os componentes conectados que contêm os vértices finais da aresta removida para obter o próximo nível hierárquico. Se o componente conectado contém pelo menos uma aresta, então, atribui-se um rótulo de novo grupo ao componente, caso contrário, atribui-se ao componente um rótulo nulo (ruído).

**end**

Fonte: adaptado de (CAMPELLO; MOULAVI; SANDER, 2013)

O grupo é considerado dividido em subgrupos menores se, e somente se, seus subgrupos forem grupos válidos à vista de alguma perspectiva (CAMPELLO; MOULAVI; SANDER, 2013; CAMPELLO et al., 2015). Já os ruídos não constituem grupos válidos, isto é, quando esse tipo de objeto é removido de um grupo, o grupo permanece com mesmo rótulo, apenas diminui de tamanho (SANTOS, 2018).

### 2.5.4 Agrupamento não-hierárquico ótimo

Os trabalhos de Campello, Moulavi e Sander (2013), Campello et al. (2015), propõem um método para obtenção de grupos significativos a partir da hierarquia gerada pelo HDBSCAN\*. Esse método é denominado *Framework for Optimal Selection of Clusters* (FOSC) e fundamenta-se no conceito de estabilidade de grupos.

A estabilidade de um grupo é fundamentada na estimativa de “sobrevivência” de um grupo ao longo da hierarquia. Sabe-se que os grupos diminuem, subdividem-se ou até mesmo desaparecem, conforme se desce na hierarquia (SANTOS, 2018). Isto posto, indica-se que os grupos mais significativos são aqueles que possuem maior sobrevida (CAMPELLO; MOULAVI; SANDER, 2013).

A medida discreta para o cálculo da estabilidade de um grupo  $C_i$ , está representada na Equação 3.

$$S(C_i) = \sum_{x_j \in C_i} (\lambda_{\max}(x_j, C_i) - \lambda_{\min}(C_i)) = \sum_{x_j \in C_i} \left( \frac{1}{\varepsilon_{\min}(x_j, C_i)} - \frac{1}{\varepsilon_{\max}(C_i)} \right) \quad (3)$$

Em que:

- $\lambda_{min}(C_i)$  representa o nível mínimo de densidade de um grupo  $C_i$ ;
- $\lambda_{max}(x_j, C_i)$  representa o nível de densidade em que o objeto  $x_j$  não pertence mais ao grupo  $C_i$ ;
- $\varepsilon_{max}(C_i)$  e  $\varepsilon_{min}(x_j, C_i)$  são os valores correspondentes ao limiar (nível de densidade)  $\varepsilon$ .

Seja  $\{C_2, \dots, C_k$  a coleção de grupos pertencentes à árvore de grupos de HDBSCAN\* (com exceção da raiz  $C_1$ ) e  $S(C_i)$  o valor de estabilidade de cada grupo, recorre-se a uma estratégia incremental e recursiva para encontrar os grupos mais significativos.

Utilizando uma abordagem *bottom-up*, a estratégia consiste em processar recursivamente cada nó da árvore e decidir qual/quais grupos devem ser melhores para a solução. Desta forma, o valor total de estabilidade dos grupos selecionados na subárvore que possui o grupo  $C_i$  é representado por  $\hat{S}(C_i)$ , conforme Equação 4.

$$\hat{S}(C_i) = \begin{cases} S(C_i), & \text{se } C_i \text{ é um nó folha} \\ \max\{S(C_i), \hat{S}(C_{il}), \hat{S}(C_{ir})\}, & \text{se } C_i \text{ é um nó interno} \end{cases} \quad (4)$$

onde  $C_{il}$  e  $C_{ir}$  correspondem aos grupos filhos esquerdo e direito, respectivamente, de  $C_i$ .

Considere-se que a árvore ilustrada na Figura 3 represente a árvore obtida pelo HDBSCAN\*, onde todos os grupos estão na solução final. Sabe-se que, se a soma das estabilidades dos grupos C8 e C9 for superior a estabilidade de C5, então, C5 é removido da solução otimizada. Assim como, se C3 apresentar uma estabilidade maior que soma das estabilidades de C6 e C7, C3 entra na solução ótima e C6 e C7 são removidos.

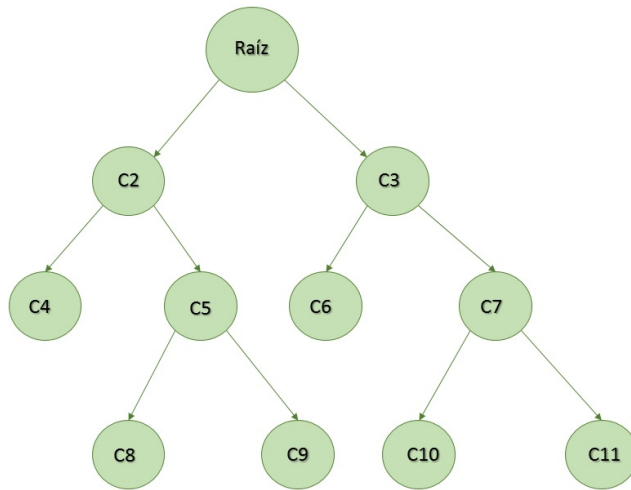


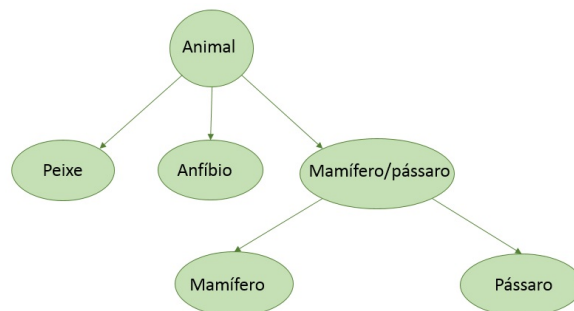
Figura 3 – Árvore de grupos - HDBSCAN\*

O FOSC indica onde cortar a árvore para obtenção dos grupos mais significativos. No entanto, foi realizado um estudo experimental do corte da hierarquia (seção 5.4 do Capítulo 5) para verificar se o FOSC realmente era uma boa opção para esta pesquisa.

## 2.6 COBWEB

O COBWEB (FISHER, 1987) é um algoritmo de agrupamento conceitual hierárquico, no qual os dados são agrupados visando maximizar as habilidades de inferência e organizados incrementalmente em uma árvore de conceitos (FERREIRA et al., 2005). Uma hierarquia conceitual é um tipo de estrutura de dados que contém um conjunto de nós parcialmente ordenados por generalidade (GENNARI; LANGLEY; FISHER, 1989), ou seja, quanto mais próximo da raiz, mais geral é o conceito, quanto mais distante, mais específico.

Figura 4 – Exemplo de uma árvore de classificação produzida pelo COBWEB



A Figura 4 ilustra uma árvore de classificação de conceitos como as que são produzidas pelo algoritmo supracitado, onde “Animal” corresponde a uma categoria mais geral e “Mamífero” representa uma categoria mais específica.

O algoritmo emprega a técnica *hill-climbing* para percorrer o espaço de possíveis sistemas de forma bidirecional buscando o melhor estado. Para isso, ele recorre a operadores aglomerativos (*merging*) e divisivos (*splitting*). Embora seja um algoritmo de agrupamento sensível à ordem de entrada dos dados, o uso dos operadores aglomerativos e divisivos representam uma estratégia para minimizar tal influência (GENNARI; LANGLEY; FISHER, 1989).

A métrica que guia a construção da árvore de conceitos é chamada de utilidade de categoria (UC), cuja fórmula está representada pela Equação 5. Essa métrica estima o ganho em se associar um objeto a um determinado grupo. A utilidade de categoria favorece agrupamentos que maximizam o potencial de inferir informações (FISHER, 1987).

$$UC = \frac{\sum_{k=1}^n P(C_k) [\sum_i \sum_j P(A_i = V_{ij} | C_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2]}{n} \quad (5)$$

Onde:  $n$  é o número de grupos em uma partição,  $C_k$  é um grupo e  $A_i = V_{ij}$  é um par atributo-valor.

A probabilidade condicional  $P(A_i = V_{ij} \mid C_k)$  indica a probabilidade do atributo  $A_i$  ter o valor  $V_{ij}$ , dado que ele pertence a um grupo  $C_k$ . Quanto maior esse valor, maior a proporção de membros do grupo compartilhando o mesmo valor. Já a probabilidade  $P(A_i = V_{ij})$  pondera a importância dos valores individuais. De forma que valores mais frequentes apresentam maior peso (ARRUDA, 2011).

Inicialmente, o CobWeb era capaz de processar apenas atributos categóricos nominais, então, no trabalho de Gennari, Langley e Fisher (1989), são propostas adaptações ao algoritmo para possibilitar o agrupamento tanto de atributos categóricos quanto de atributos numéricos. Essa versão melhorada é que será adotada nesta pesquisa.

A métrica de avaliação revisada, capaz de lidar com dados numéricos, está representada na Equação 6

$$UC_{num} = \frac{\sum_{k=1}^n P(C_k) \sum_{i=1}^m \frac{1}{\sigma_{ik}} - \sum_{i=1}^m \frac{1}{\sigma_{ip}}}{n} \quad (6)$$

Onde:  $m$  é o número de atributos,  $n$  é a quantidade de grupos,  $\sigma_{ik}$  é o desvio-padrão de determinado atributo em relação a determinado grupo e  $\sigma_{ip}$  é o desvio-padrão de determinado atributo em relação ao nó pai.

O pseudocódigo representado pelo Algoritmo 2 permite observar o funcionamento do COBWEB. Para cada objeto a ser inserido, a árvore de classificação é percorrida, começando pela raiz até as folhas, e um dos seguintes operadores - o que apresentar a maior UC - é aplicado: incorporar a um nó existente, criar novo nó, *merge* ou *split*.

Embora seja uma opção interessante para o problemática levantada neste projeto, o trabalho de Duarte (2008) aponta algumas limitações do COBWEB, tais como:

- ❑ O algoritmo assume que os atributos são estatisticamente independentes uns dos outros, o que nem sempre é verdade;
- ❑ O custo do cálculo da distribuição de probabilidades aumenta quando os atributos possuem um grande número de valores possíveis.
- ❑ A árvore de classificação não é balanceada em altura, ou seja, as alturas das sub-árvores podem diferir em mais de um nível. Dessa forma, a complexidade temporal e espacial pode se degradar drasticamente em grandes conjuntos de dados.

**Algoritmo 2** COBWEB**Entrada:** O nó atual  $N$  da hierarquia de conceito e a nova instância  $I$ .**Saída:** Hierarquia de conceito que classifica a instância  $I$ .Variáveis:  $N$  é o nó atual; $I$  é uma instância não classificada; $C$ ,  $P$ ,  $Q$  e  $R$  são nós na hierarquia; $U$ ,  $V$ ,  $W$  e  $X$  são as utilidades de categoria (UC) dos grupos.**begin**  **if**  $N$  é um nó terminal **then**    | Cria um novo nó terminal( $N, I$ ) e insere( $N, I$ )  **end**  **else**    Incorpora ( $N, I$ )    **for** cada filho  $C$  do nó  $N$  **do**      | calcular a UC ao inserir a instância  $I$  em  $C$       | Seja  $P$  o nó que obtém a maior utilidade de categoria  $W$  ao incorporar  $I$ .      | Seja  $R$  o nó que obtém a segunda maior utilidade de categoria  $U$  ao incorporar  $I$ .      | Seja  $X$  a utilidade de categoria de incorporar a instância  $I$  a um novo nó  $Q$ .      | Seja  $Y$  a utilidade de categoria de fazer *merging* dos nós  $P$  e  $R$ .      | Seja  $Z$  a utilidade de categoria de fazer *splitting* em  $P$ .      **if**  $W$  é a maior UC **then**        | Incorpora  $I$  em  $P$  e chama Cobweb( $P, I$ ).      **end**      **if**  $X$  é a maior UC **then**        | Cria um novo nó  $Q$  e inicializa as probabilidades com relação a  $I$       **end**      **if**  $Y$  é a maior UC **then**        | A variável  $O$  recebe Merge( $P, R, N$ ) e chama Cobweb( $O, I$ )      **end**      **if**  $Z$  é a maior UC **then**        | A variável  $O$  recebe Split( $P, N$ ) e chama Cobweb( $N, I$ )      **end**    **end**  **end****end**

Fonte: adaptado de (GENNARI; LANGLEY; FISHER, 1989)

## 2.7 Comparação de hierarquias

No trabalho de Neto et al. (2019), foi proposto um método chamado *Hierarchy Agreement Index* (HAI), capaz de comparar o quanto duas hierarquias são semelhantes entre si. Essa medida pode ser aplicada com a finalidade de mensurar se a variação dos parâmetros de um algoritmo de agrupamento hierárquico implica hierarquias diferentes ou, ainda, comparar hierarquias produzidas quando se varia ordem de entrada dos dados.

A distância entre um par de pontos  $x_i$  e  $x_j$  em uma hierarquia  $H$ , representada por



$d_H(x_i, x_j)$ , é definida como o tamanho do menor grupo em que  $x_i$  e  $x_j$  aparecem juntos, divididos pelo tamanho do conjunto de dados inteiro. A semelhança HAI entre duas hierarquias  $H_1$  e  $H_2$  é, então, definida pela média normalizada da diferença das distâncias  $d_{H_1}$  e  $d_{H_2}$ , entre todos os pares de pontos, conforme a Equação 7.

$$HAI(H_1, H_2) = 1 - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |d_{H_1}(x_i, x_j) - d_{H_2}(x_i, x_j)| \quad (7)$$

Onde o valor de HAI pertence ao intervalo  $[0, 1]$ , sendo que quanto mais próximo de 1, mais semelhantes são as hierarquias comparadas.

## 2.8 Medidas de validação

Segundo Tan et al. (2006), algoritmos de agrupamento podem produzir soluções mesmo quando são aplicados à base de dados que não possuem nenhuma estrutura. Portanto, para avaliar se o agrupamento encontrado por determinado algoritmo representa uma boa solução, faz-se necessário avaliar os resultados por meio de medidas de validação.

Os índices ou medidas de validação são divididos em três categorias, a saber: externos, internos e relativos.

As medidas de validação baseadas em critérios externos medem o grau de proximidade entre o agrupamento encontrado por determinado algoritmo e um agrupamento de referência (tido como ideal), o que exige conhecimento prévio da estrutura da base de dados (HORTA, 2013). Como exemplo de índice externo, pode-se assinalar o *Rand Index* (RAND, 1971), representado pela Equação 8, no qual  $U$  representa o agrupamento avaliado e  $V$  representa o agrupamento ideal.

$$RI(U, V) = \frac{n_{00} + n_{11}}{n_{00} + n_{10} + n_{01} + n_{11}} \quad (8)$$

As demais variáveis são definidas da seguinte forma (HORTA, 2013):

- $n_{00}$ : número de pares de instâncias contidas nos mesmos grupos, tanto em  $U$  quanto em  $V$ ;
- $n_{10}$ : número de pares de instâncias contidas em grupos diferentes em  $U$ , mas nos mesmos grupos em  $V$ ;
- $n_{01}$ : número de pares de instâncias contidas em grupos diferentes em  $V$ , mas nos mesmos grupos em  $U$ ;
- $n_{11}$ : número de pares de instâncias contidas em grupos diferentes, tanto em  $U$  quanto em  $V$ ;

Entrementes, os índices internos qualificam o agrupamento pautando-se apenas em informações intrínsecas ao próprio conjunto de dados (TAN et al., 2006), sem conhecimento prévio sobre a estrutura dos dados. Os índices relativos são índices internos que comparam quantitativamente a diferença entre os valores dos índices de validação de dois agrupamentos gerados (NALDI, 2011).

Evidencia-se o Coeficiente de Silhueta (ROUSSEEUW, 1987) como exemplo de índice de validação interno. Ele é calculado para cada objeto de um grupo, a fim de verificar se ele está bem alocado em seu grupo ou se deveria estar em outro grupo. Para realizar tal cálculo, qualquer medida de similaridade/dissimilaridade pode ser adotada (NALDI, 2011).

Dado um objeto  $x_i$  pertencente a um grupo  $C_A$ , seja  $a(x_i)$  a distância média do objeto  $x_i$  em relação a todos os objetos do grupo  $C_A$  e  $b(x_i)$  a menor distância em relação a um grupo  $C_B$ , onde  $C_B \neq C_A$ . A fórmula geral do Coeficiente de Silhueta está indicado na Equação 11.

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max[a(x_i), b(x_i)]} \quad (9)$$

O resultado obtido pela Equação 11 pertencerá ao intervalo  $[-1, 1]$ , sendo que quanto mais próximo de 1, melhor. Por fim, para se obter a silhueta do agrupamento, deve-se somar os Coeficientes de Silhueta de todos os objetos do conjunto de dados e dividir pelo número total de objetos do conjunto, representado por  $n$ , conforme Equação 10.

$$Silhueta = \sum_{i=1}^n \frac{S(x_i)}{n} \quad (10)$$

O Coeficiente de Silhueta utiliza a distância média dos objetos em relação a todos os objetos do mesmo grupo e a menor distância em relação a um objeto de outro grupo, fazendo-se necessário calcular a distância de todos contra todos, fator que implica alto custo computacional.

Uma desvantagem do Coeficiente de Silhueta é justamente esse alto custo computacional, cuja complexidade temporal é da ordem de  $O(n^2)$  (ALVES, 2007). Para tanto, uma solução apresentada em Vendramin, Campello e Hruschka (2010) é o uso da silhueta simplificada.

Uma alternativa para diminuir esse custo computacional é adotar a Silhueta Simplificada, uma vez que essa emprega a distância dos objetos ao centroide do grupo ao qual pertence e a distância ao centroide do grupo mais próximo. Isso reduz a complexidade temporal para  $O(n)$  (ALVES, 2007).

## 2.9 Considerações Finais

Neste capítulo, foram abordados os conceitos fundamentais para a compreensão deste trabalho. Primeiro, foi apresentada uma visão geral da Comunicação de Acidentes de Trabalho e quais medidas o Ministério Público tem tomado para prevenir a ocorrência de acidentes. Em seguida, foi descrito todo o processo de Descoberta de Conhecimento em Bases de Dados, dando o devido destaque para a tarefa de agrupamento, que será aplicada na fase experimental deste trabalho. Por fim, foram expostos os conceitos e definições dos algoritmos HDBSCAN\* e CobWeb, bem como uma forma de comparação de hierarquia e a medida de validação que serão empregados neste estudo.

No Capítulo 3, serão exibidos os trabalhos relacionados e o modo como este estudo confronta ou se associa a cada um deles.

---

## Trabalhos relacionados

Neste capítulo, serão apresentados os trabalhos relacionados a este estudo. A Seção 3.1 retrata a abordagem das ciências humanas, na qual não há a aplicação de técnicas de extração de conhecimento, mas expõe algumas consequências e impactos gerados pelos acidentes de trabalho. Enquanto a Seção 3.2 explana sobre os trabalhos desenvolvidos com o apoio de técnicas de extração de conhecimento aplicadas a estudos acerca da temática desta pesquisa, e destaca suas principais diferenças quanto a esta. Por fim, a Seção 3.3 discorre sobre as considerações finais deste capítulo.

### 3.1 Abordagem das ciências humanas

Nesta seção, serão expostos alguns trabalhos que, embora não utilizem técnicas computacionais para a extração de conhecimento, abordam a temática relativa aos acidentes de trabalho.

No trabalho desenvolvido por Almeida e Barbosa-Branco (2011), foi proposto um estudo com a finalidade de estimar a duração e o custo dos benefícios previdenciários pagos, em decorrência de acidentes de trabalho, pelo Instituto Nacional de Seguridade Social (INSS) no ano de 2008. Os autores estratificaram as informações quanto ao sexo do segurado, idade e Classificação Internacional de Doenças (CID).

Os autores observaram que os acidentes mais recorrentes foram os classificados como: i) Lesões, envenenamento e algumas outras consequências de causas externas; ii) Doenças do sistema osteomuscular e do tecido conjuntivo; e iii) Transtornos mentais e comportamentais. O primeiro grupo representa a maioria dos casos, sugerindo precariedade das medidas de segurança no trabalho. Além disso, constatou-se que a prevalência dos benefícios foi maior no sexo masculino na faixa etária maior ou igual a 40 anos. E homens e mulheres apresentam diferentes perfis de acidentabilidade, sem interferência na duração do benefício.

O estudo desenvolvido por Torres et al. (2011) teve como objetivo identificar as repercussões do adoecimento no contexto familiar do trabalhador com lesões por esforços

repetitivos (LER) ou distúrbios osteomusculares relacionados ao trabalho (DORT), e descrever as estratégias usadas para o enfrentamento de doenças relacionadas ao trabalho.

As conclusões obtidas pelo trabalho de Torres et al. (2011) indicaram que alterações psicoafetivas, sinais de comprometimento na relação familiar, perda da autoestima e isolamento social, e os mecanismos para amenizar as angústias frente ao adoecimento vão desde as atividades de lazer à reabilitação física. Os autores ressaltam, ainda, a importância de reforçar ações de prevenção, vigilância e atenção, visando à redução de danos ocasionados pelos perigos advindos dos processos produtivos.

O estudo de Santana et al. (2006) focou em mensurar a proporção de benefícios concedidos por acidentes de trabalho dentre o total de benefícios relacionados com a saúde da Previdência Social, e em analisar o impacto sobre a produtividade relativa aos dias de trabalho que foram perdidos. Para isso, foram analisados 31.096 benefícios concedidos por doenças ou agravos à saúde, dos quais 2.857 eram devido a acidentes de trabalho. Os custos com os benefícios por acidentes de trabalho foram estimados em R\$8,5 milhões, com, aproximadamente, meio milhão de dias perdidos de trabalho no ano.

Segundo Santana et al. (2006), apesar do fato de que esses dados são sub-enumerados, e restritos aos trabalhadores que conseguiram receber benefícios relativos à saúde, os resultados demonstram o grande impacto sobre a produtividade e o orçamento do Instituto Nacional de Previdência Social de agravos reconhecidos como evitáveis, reforçando a necessidade de sua prevenção.

Os trabalhos apontados nesta seção têm como foco estimar os custos dos benefícios previdenciários pagos em decorrência dos acidentes de trabalho e analisar os impactos gerados no aspecto social e emocional do trabalhador, sem a aplicação de técnicas de extração de conhecimento. Tal característica difere do objetivo desta dissertação, que têm como foco a busca por uma estrutura de grupos nos dados do MPT, mediante a aplicação de técnicas de extração de conhecimento.

## **3.2 Abordagem usando técnicas de extração de conhecimento**

Voltando o foco para os trabalhos que empregam técnicas de extração de conhecimento, destacam-se os trabalhos de Bartolomeu (2002), Porto e Júnior (2006), Silva (2018), Brito (2019) e Rodrigues (2019). As principais características desses trabalhos estão sumarizadas na Tabela 3. Por fim, serão expostas algumas notáveis descobertas expostas na plataforma do Observatório Digital de Segurança e Saúde no Trabalho.

Tabela 3 – Trabalhos relacionados

Autor(es)	Base de dados	Técnica empregada
Bartolomeu (2002)	Acidentes de trabalho ocorridos no estado de Santa Catarina em 2000	Distribuição de frequência, teste de hipótese e correlação de variáveis
Porto e Júnior (2006)	Acidentes de trabalho dentro de um hospital específico	Algoritmo de agrupamento: <i>k-means</i>
Silva (2018)	Benefícios Previdenciários (SISBEN)	Algoritmos de agrupamento: <i>k-means</i> , <i>Canopy</i> e <i>Expectation Maximization</i> (EM)
Brito (2019)	Acidentes de trabalho (CATWEB)	Visualização de dados: <i>Dendogram Composition</i> e <i>Choropleth Composition</i>
Rodrigues (2019)	Acidentes de trabalho (CATWEB)	Visualização de dados: projeções multidimensionais e <i>layouts</i> hierárquicos

Em Bartolomeu (2002), foi proposto um modelo para investigação dos dados sobre acidentes de trabalho e doenças ocupacionais, baseado na extração de conhecimento mediante grande volume de dados.

Para fins experimentais, os autores valeram-se de acidentes notificados ao INSS, no estado de Santa Catarina, no ano 2000. O modelo proposto é constituído pelas seguintes etapas:

1. Definição da estratégia para obtenção dos dados: delimitação do problema investigado, identificação das variáveis relevantes ao estudo, a forma de obtenção dos dados e a migração dos dados para a ferramenta computacional utilizada.
2. Exploração dos dados brutos: estudo dos dados, a fim de conhecer seus tipos (texto, data, número, dentre outros) e os domínios (valores possíveis para cada variável).
3. Tratamento dos dados: verificação e/ou criação de codificação para os valores de domínio de uma variável, estruturação dos dados, eliminação de erros e inconsistências, padronização dos dados e agrupamento dos domínios das variáveis (Exemplo: agrupar a variável "idade" por faixas etárias).
4. Análise estatística dos dados: delineamento do perfil das CATs registradas, das empresas empregadoras, dos trabalhadores acidentados, dos acidentes e doenças ocorridas, utilizando técnicas como distribuição de frequência, teste de hipótese e correlação de variáveis.
5. Descoberta do conhecimento: aplicação de técnicas de *data mining* para encontrar padrões. Foram utilizadas duas abordagens para a extração de conhecimento,

com e sem formulação prévia de hipóteses. Os autores utilizaram uma ferramenta denominada "See5 para a geração das regras.

6. Ranqueamento das regras geradas: Organização das regras, obtidas na etapa anterior, seguindo algum critério.
7. Análise e avaliação dos conhecimentos descobertos: verificação do valor dos conhecimentos gerados.
8. Simulação do impacto da adoção de ações: simulação de ações gerenciais e/ou estratégicas que possam ser tomadas valendo-se dos conhecimentos obtidos.

O modelo foi considerado eficiente para a proposta, tornando possível identificar correlações, padrões, informações implícitas e regras que caracterizam tendências em um curto espaço de tempo. Os autores destacaram o alto índice de campos deixados em branco no preenchimento das CATs, indicando uma possível dificuldade dos usuários para o preenchimento dos formulários.

A presente pesquisa diferencia-se do trabalho de Bartolomeu (2002) quanto à técnica aplicada para extração de conhecimento e quanto ao conjunto de dados analisado. Enquanto esse recorre a técnicas como distribuição de frequência, teste de hipótese e correlação de variáveis, e restringe-se ao estado de Santa Catarina no ano 2000, o trabalho de dissertação aqui proposto recorre à tarefa de agrupamento, a fim de identificar padrões nos dados do MPT, e utiliza, como base de dados, as CAT's registradas em todo o Brasil, referentes aos anos de 2012 a 2017.

O estudo de Porto e Júnior (2006) utilizou uma técnica de agrupamento particional, o algoritmo k-means, a fim de identificar grupos com características em comum, em relação aos acidentes do trabalho ocorridos em determinado hospital. O objetivo da pesquisa foi encontrar, mensurar e descrever as causas e respectivos custos incorridos pela empresa, possibilitando uma gestão mais efetiva dos valores gastos.

A análise de dados foi realizada por meio de um software estatístico chamado *Statistical Package for the Social Sciences* (SPSS), e não foi aplicada nenhuma medida de validação para avaliação dos agrupamentos obtidos. Por meio de análises de clusters, foi identificada a formação de quatro grupos:

- ❑ Cluster 01 - Acidentes com Auxiliares de Enfermagem lotados nas Unidades de Terapia Intensiva (UTI's) e Internação, devido a cortes, contusões e perfurações.
- ❑ Cluster 02 - Acidentes com Auxiliares de enfermagem lotados nas UTIS, Internação e Centro Cirúrgico.
- ❑ Cluster 03 - Acidentes com Auxiliares de enfermagem lotados nas UTIS e Internação devido a perfurações.

- Cluster 04 - Acidentes com Auxiliares de Farmácia, Higiene e Lavanderia lotados na Lavanderia, Higiene e Farmácia.

Os autores ressaltam que os setores de maior frequência de acidentes de trabalho são: Internação, Centro Cirúrgico, Higiene e Limpeza. O cargo mais susceptível é o de Auxiliar de Enfermagem. Dos acidentes registrados, verificou-se que o agente causador de maior incidência de acidente foi a agulha.

O total de gastos adicionais para o hospital, decorrentes dos acidentes sofridos no período de 2003 a 2005, foi de R\$74.121,90 e são provenientes de atendimento médico ambulatorial, transportes, medicações, custos com exames laboratoriais e/ou imagens e custos com reposição. Os autores afirmam que, por meio da contabilidade de custos e da análise de agrupamentos, é possível identificar, mensurar e descrever as causas e respectivos custos incorridos pela empresa, o que pode permitir uma gestão mais efetiva desses valores.

Destaca-se que o trabalho de Porto e Júnior (2006) utiliza o algoritmo de agrupamento *k-means*, algoritmo baseado no método particional, e explora apenas os dados referentes a determinado hospital. Enquanto a pesquisa proposta por esta dissertação recorre a algoritmos de agrupamento baseados no método hierárquico divisivo (HDBSCAN\*) e conceitual hierárquicos (COBWEB), e é aplicado a todos os tipos de acidentes de trabalho registrados por meio de CAT's.

Outrossim, o objetivo de Porto e Júnior (2006) foi encontrar as causas mais frequentes dos acidentes de trabalho ocorridos no hospital e mensurar os gastos, possibilitando uma gestão mais efetiva. Ao passo que o objetivo aqui estabelecido é buscar padrões nos dados relativos aos acidentes de trabalho, visando fornecer informações que possam contribuir com a criação de políticas públicas para a prevenção de acidentes e/ou outras tomadas de decisões pelos gestores públicos.

No projeto desenvolvido por Silva (2018), seu Trabalho de Conclusão de Curso (TCC), os algoritmos de agrupamento *K-means*, *Canopy* e *Expectation Maximization (EM)* foram aplicados à base de dados de Benefícios Previdenciários (SISBEN) do Ministério Público do Trabalho. Os resultados experimentais foram avaliados por meio da medida de validação denominada Silhueta Simplificada. Diferentes versões da base foram utilizadas nos experimentos. No entanto, após avaliação por medida de validação e interpretação do especialista de domínio, constatou-se que não foram alcançados resultados satisfatórios. Uma provável justificativa para isso foi o fato de que, antes de aplicar os algoritmos de agrupamento, foi necessário converter os atributos categóricos em numéricos, o que gerou uma base de alta dimensionalidade, esse fator prejudicou o desempenho de tais algoritmos.

Contudo Silva (2018) deixou como contribuição uma ferramenta para pré-processamento de dados e sugere, ainda, como trabalhos futuros, o uso de métodos de agrupamento mais complexos e que explorem outras vertentes, tais como: agrupamento relacional, agrupamento por densidade, agrupamento em *grid* e hierárquico.



Diferentemente do trabalho de Silva (2018), a base de dados utilizada nesta dissertação é a de acidentes de trabalho (CATWEB) e os algoritmos de agrupamento de dados aplicados são mais complexos e baseados em métodos diferentes daqueles que foram propostos pelo autor supracitado.

O trabalho de Brito (2019) valeu-se de uma estratégia de exploração visual para realizar uma análise dos aspectos temporais e geográficos dos acidentes de trabalho. Os *layouts* temporais foram projetados para revelar a evolução dos acidentes ao longo do tempo, como comportamento anômalo e sazonalidade, entre outras situações. Enquanto que os *layouts* geográficos possibilitam uma navegação entre as localidades, explorando o contexto e comparando as particularidades de cada localidade.

Para a visualização dos dados, a autora recorreu ao *Dendogram Composition* e ao *Choropleth Composition*. O *Dendogram* fornece comparações em vários níveis da evolução dos acidentes de trabalho no Brasil durante um único ano, e o *Choropleth* fornece contexto geográfico e comparação da evolução dos acidentes de trabalho dentro de uma localidade específica escolhida em vários anos. Após os experimentos, as seguintes conclusões foram destacadas:

- ❑ Em 2012, a região Sudeste registrou o maior número de acidentes de trabalho, seguido pela região Sul. Enquanto a Região Norte registrou o menor número de acidentes. Embora as regiões Norte e Nordeste apresentem um número menor de acidentes de trabalho registrados, sabe-se que tais regiões têm alguns problemas estruturais, como: condições de trabalho precárias e trabalhadores informais. Portanto, os padrões verificados sugerem que os acidentes de trabalho podem ter sido subnotificados.
- ❑ Na região Nordeste, ainda no ano de 2012, observaram-se picos de onda: dois relativos à ingestão de comida, que acometeu mais de 200 trabalhadores, em dois dias específicos, e o outro relativo a um acidente de trânsito, que culminou em 208 registros. No entanto, tais picos representam um padrão de evento localizado, o que não influencia a análise visual dos dados da região.
- ❑ O estado de Alagoas registrou, em 2012, o maior número de acidentes de trabalho, sendo um total de 5.647, e uma redução gradual pôde ser observada ao longo dos anos. De tal forma que, em 2017 a redução foi de 50,4% em relação a 2012, totalizando 2801 acidentes registrados. Os acidentes ocorrem em sua maioria, no início e no fim do ano, e mais concentrado na mesorregião Leste Alagoano, e estão relacionados às duas maiores atividades econômicas nessa área: mineração de ouro e abate de animais.
- ❑ Analisando cidades distantes, como Ananindeua, localizada na região Norte, e Paranaguá, na região Sul, percebeu-se que, embora Ananindeua tenha uma população

3 vezes maior que Paranaguá, foram registrados 35% menos acidentes nesta do que naquela. Cidades pequenas que relatam um alto número de acidentes ou grandes cidades que relatam poucos acidentes, são situações que merecem atenção e uma investigação mais profunda, pois os acidentes podem não estar sendo adequadamente registrados.

- ❑ Observou-se, ainda, uma redução na proporção de acidentes registrados ao longo dos anos, em todo país. Sugerindo que melhores medidas de segurança podem ter sido tomadas. Outro fato observado é que o número de acidentes reduz no mês de Dezembro, o que pode ser explicado pelo grande número de trabalhadores que tiram férias nesse período.
- ❑ Por fim, no estado de Minas Gerais, duas mesorregiões se destacam quanto a quantidade de acidentes registrados: Metropolitana de Belo Horizonte e Triângulo Mineiro/Alto Paranaíba. Embora a primeira contenha a capital do estado, é na segunda mesorregião que ocorre a maior incidência de acidentes.

Seguindo na linha de exploração visual, Rodrigues (2019) também propõe uma análise estrutural visual dos dados de acidentes de trabalho. O autor valeu-se de projeções multidimensionais e *layouts* hierárquicos para compreender questões como a influência de cada atributo na caracterização dos grupos visualizados, a identificação de correlações entre os atributos envolvidos, o comportamento de tendências, entre outras tarefas.

No trabalho em questão, também foram consultadas fontes externas, como notícias e dados brutos, para corroborar os *insights* obtidos com a visualização dos dados. Dentre os achados, vale destacar:

- ❑ Ao associar elementos como atividade econômica e localização geográfica, o autor verificou que duas atividades sobressaem como as mais significantes na maioria das cidades, são elas: "Indústrias de transformação" e "Comércio/reparação de veículos".
- ❑ Ao observar um conjunto de cidades de "interesse particular", verificou-se que a maioria de seus acidentes ocorreram na área de "Serviços sociais e saúde humana". Esse conjunto é composto por 127 cidades, incluindo capitais e grandes cidades. Cada uma dessas cidades registrou em média 3.000 acidentes nessa atividade econômica, valor bem mais alto que em qualquer outra categoria. Tal análise demonstrou como a atividade econômica e o número de acidentes são diferentes em cidades grandes/ricas, em relação ao resto do país, independentemente da região a qual pertençam.
- ❑ Em 4 das 5 regiões do país, o agente causador "Máquinas e equipamentos" foi o que apresentou maior número de CAT's registrados. A exceção é a região Norte, apesar de esse agente causador ainda estar no topo da lista dos mais frequentes.

- ❑ Os membros superiores constituem as partes do corpo atingida mais afetada, exceto quando se trata de áreas rurais.
- ❑ Brasília e Santo André são cidades com comportamentos semelhantes entre si, em relação à ocorrência de acidentes de trabalho. E comportamentos distintos das regiões às quais pertencem, Centro-Oeste e Sudeste, respectivamente.
- ❑ Analisando Brasília, observou-se que a distribuição de acidentes por sexo é mais equilibrada, ainda que o maior percentual seja masculino (62%). Além disso, as ocupações "Arte e Ciência e "Saúde Humana têm mais acidentes envolvendo mulheres, enquanto "Produção de bens e serviços têm mais acidentes com pessoas do sexo masculino. A ocupação com o maior número de CAT's registradas é "Prestadores e vendedores de serviços.
- ❑ Nas regiões Nordeste e Centro-Oeste, ocorreu uma quantidade significativa de acidentes na área rural, cerca de 8%. O percentual de acidentes envolvendo pessoas do sexo masculino supera 70%. E os agentes causadores mais recorrentes são: "Máquinas e equipamentos", "Agente químico", "Agente Biológico" e "Veículos".
- ❑ Na região Norte, 80% dos acidentes registrados foram de pessoas do sexo masculino. Embora "Indústrias de transformação ainda configure a atividade econômica com maior número de CAT's, a segunda maior atividade é "Construção", comportamento que não se repete nas demais regiões do país. O agente causador mais reportado é "Agente químico".
- ❑ O estado do Pará e Amazonas são responsáveis por 70% do número de acidentes registrados na região Norte. Sendo que, em suas capitais, os acidentes estão mais associados à Indústria, enquanto, no resto do estado, os acidentes estão ligados à Pecuária e Agricultura.

Embora tanto o trabalho de Brito (2019) quanto o de Rodrigues (2019) explorem a mesma base de dados (CATWEB), os autores valem-se de visualização de dados para extração de conhecimento. Enquanto, nesta pesquisa, serão utilizadas técnicas de agrupamento de dados.

Além disso, Brito (2019) focou nos aspectos temporais e geográficos dos acidentes de trabalho, e Rodrigues (2019) verificou a influência dos atributos na caracterização dos grupos visualizados. Já nesta dissertação, será realizado um estudo dos atributos, a fim de verificar quais contribuem para a tarefa de agrupamento, contribuindo para a descoberta de padrões que possam ser interessantes ao Ministério Público do Trabalho.

O Observatório Digital (Seção 2.2.1) possui uma aba denominada "Achados", em que são apresentadas algumas descobertas sobre os dados de acidentes de trabalho e benefícios previdenciários, como gráficos com informações sobre as atividades econômicas em que

mais ocorrem acidentes, as partes do corpo mais frequentemente atingidas, o perfil etário e de sexo em relação ao número de registros de acidentes, dentre outros.

Os painéis, gráficos e visualizações disponíveis no Observatório Digital, permitem relacionar apenas dois atributos de cada vez. Por exemplo: é possível verificar a distribuição geográfica dos acidentes de trabalho relacionando a quantidade de CAT's registradas em relação aos Municípios ou UF's, ou ainda, qual o tipo de lesão mais frequente relacionando a quantidade de CAT's com a natureza da lesão.

Em suma, percebe-se uma certa limitação quanto à exploração dos dados, sendo possível relacionar apenas dois atributos de cada vez. Mediante isso, esta pesquisa surge com o propósito de explorar mais atributos ao mesmo tempo, e as relações entre os atributos serão obtidas de forma intrínseca pelos algoritmos de agrupamento.

### 3.3 Considerações finais

Neste capítulo, foram destacados trabalhos encontrados na literatura que se relacionam com a temática da pesquisa aqui desenvolvida. Além disso, estabeleceu-se uma relação e diferenciação deles para com este, destacando qual a inovação que justifica esta pesquisa.

O Capítulo 4 detalhará a proposta desenvolvida, discorrendo sobre a formalização do problema, a apresentação da base de dados, a descrição dos pré-processamentos realizados e demais procedimentos necessários para alcançar o objetivo almejado.

---

## Método para detecção e análise de grupos na CATWEB

A OIT aponta que 4% do Produto Interno Bruto são perdidos em razão de acidentes de trabalho e doenças ocupacionais (FUNDACENTRO, 2019). Além do alto custo para o INSS, deve-se considerar os demais prejuízos causados aos trabalhadores, aos familiares e ao empregador, tais como: perda de produtividade, alterações psicoafetivas, sentimento de medo e até a perda de entes queridos. Diante dessa realidade, faz-se notável a necessidade da criação de políticas públicas direcionadas para evitar a ocorrência de acidentes de trabalho e para a prevenção de doenças ocupacionais.

Valendo-se da base de dados CATWEB (detalhada na Subseção 4.1) e empregando o método de Descoberta de Conhecimento em Bases de Dados (apresentado na Seção 2.3), este trabalho busca caracterizar grupos e conferir interpretabilidade aos resultados obtidos, a fim de extrair informações. Mediante o uso de algoritmos de agrupamento, será possível estabelecer relações de semelhança entre os acidentes de trabalho (acidentes pertencentes ao mesmo grupo) e diferenças (acidentes pertencentes a grupos distintos). Ademais, caracterizando e interpretando os maiores grupos, será possível fornecer informações relevantes ao Ministério Público do Trabalho, a fim de apoiar o processo de tomadas de decisões para prevenção dos acidentes de trabalho e atenuação dos infortúnios acarretados.

Neste capítulo, será descrito o método adotado para detecção e análise de grupos da base de dados CATWEB. A Seção 4.2 descreve todas as etapas do método e como foram executadas. A Seção 4.3 relata as adaptações realizadas para a utilização dos algoritmos HDBSCAN\* e Cobweb. Por fim, a Seção 4.4 tece as considerações finais.

### 4.1 Base de dados

No Observatório Digital de Saúde e Segurança do Trabalho, estão disponíveis dois conjuntos de dados: os Dados de Acidentes de Trabalho Notificados (CATWEB) e os

Dados de Benefícios Previdenciários (SISBEN), referentes aos anos de 2012 à 2017.

A CATWEB foi definida como objeto de estudo desta dissertação, pois, para reduzir o número de acidentados e o gasto com benefícios acidentários, é preciso, primeiramente, conhecer melhor os acidentes de trabalho.

No entanto, foram percebidos alguns fatores que configuram verdadeiros desafios para a análise desses dados:

1. O tamanho da base de dados, visto que a CATWEB é composta por 3.879.755 de instâncias;
2. O fato de que 14 dos 18 atributos da base são categóricos, sendo que alguns atributos possuem um grande número de categorias possíveis. O atributo “Município”, por exemplo, possui um total de 5.285 valores distintos;
3. Grande quantidade de valores ausentes;
4. Diversidade de acidentes de trabalho;
5. Grande diversidade das regiões do país, o que pode ocasionar diferentes tipos de acidentes de trabalho

Além disso, destaca-se que a CATWEB não conta com conhecimento prévio sobre a organização dos dados ou rótulos de classes (*ground truth*). Diante desses desafios, foi delineado um método de trabalho, descrito na 4.2, que contempla o pré-processamento da base, a escolha dos algoritmos usados, bem como as adaptações feitas nestes, e a forma de avaliação.

A seguir, todos os atributos presentes na CATWEB serão descritos em detalhes e, para facilitar a visualização das principais características, o resumo destas informações segue ilustrado na tabela 4.

1. **st\_acidente\_feriado**: o indicador de acidente em feriado é um atributo categórico nominal, composto por nove valores possíveis. São eles: Carnaval, Confraternização Universal, *Corpus Christi*, Dia do Trabalho, Finados, Independência do Brasil, Natal, Nossa Senhora Aparecida (Padroeira do Brasil) e Paixão de Cristo. Trata-se de um atributo com uma quantidade considerável de valores ausentes, sendo que 3.827.313 instâncias (98,65% da base de dados) não apresentam esse campo preenchido. Esse atributo somente possui seu valor preenchido, quando o acidente ocorreu em feriado. No entanto, quando o valor não se encontra preenchido, existem duas possibilidades: o acidente não ocorreu em feriado ou o atributo não foi devidamente preenchido.
2. **ds\_agente\_causador**: o agente causador pode ser uma ferramenta, substância química ou física, objeto ou simplesmente uma causa natural ou não, que ocasionou

- o acidente de trabalho. Trata-se de um atributo categórico nominal, com 302 valores possíveis. O número de valores ausentes corresponde a 861.008 instâncias (22,19% do total).
3. **ano\_cat**: o ano do acidente é um atributo do tipo numérico intervalar, que varia entre 2012 e 2017. Não há valores ausentes para esse atributo.
  4. **ds\_cnae\_classe\_cat**: a Classificação Nacional de Atividade Econômica (CNAE) (CONCLA, 2019) é uma forma de padronizar os códigos de atividade econômica em todo o país. Esse atributo possui um total de 668 valores distintos, é do tipo categórico nominal. Nesse conjunto de dados, o número de valores ausentes é de 46.038 (1,19% do total da base) e 292.277 receberam o valor “Indefinido” (7,53%).
  5. **dt\_acidente**: a data do acidente é um atributo numérico intervalar. Está disposto da seguinte forma: “DD/MM/AAAA”, onde DD representa o dia do mês, MM representa o mês e AAAA representa o ano. Não há valores ausentes para esse atributo.
  6. **st\_dia\_semana\_acidente**: o dia da semana é um atributo categórico ordinal, com sete valores possíveis: segunda, terça, quarta, quinta, sexta, sábado e domingo. Não há valores ausentes para esse atributo.
  7. **ds\_emitente\_cat**: o emitente é quem se responsabilizou por registrar a CAT, podendo ser o próprio acidentado, uma autoridade pública, o empregador, o médico ou o sindicato. Trata-se de um atributo categórico nominal. O número de valores ausentes corresponde a 852.877 instâncias (21,98% do total). Os especialistas do MPT destacaram que uma possibilidade para o campo aparecer em branco é o preenchimento automático da CAT por um sistema integrado. Isso acontece quando um benefício é fornecido ao acidentado, ou seja, ele aparece como uma instância na base de benefícios do INSS, mas não foi registrado o acidente na base de CATs. Nesse caso, vários outros campos da CAT poderão ficar incompletos
  8. **hora\_acidente**: a hora em que ocorreu o acidente é um atributo numérico intervalar. Embora não haja valor ausente, os especialistas de domínio informaram que não há garantias de que o valor preenchido corresponda, de fato, ao horário do acidente.
  9. **idade\_cat**: a idade do acidentado é um atributo numérico racional. Para esse conjunto de dados, os valores possíveis variam de 0 a 98 anos. Na CATWEB, foram verificadas 117 instâncias cujo valor para o atributo “idade” é inferior a 16 anos. No Brasil, o trabalho é proibido para quem ainda não completou 16 anos, como regra geral. Dessa forma, considerou-se que os dados não foram corretamente preenchidos ou trata-se de trabalho infantil, o que é ilegal. Portanto, essas instâncias foram desconsideradas. Apenas 173 instâncias não foram preenchidas com nenhum valor.

10. **cd\_indica\_obito**: o indicador de óbito é um atributo binário, que informa se o acidente de trabalho teve como consequência a morte do acidentado. Um total de 852.862 instâncias não tiveram o valor preenchido para esse atributo (21,98% do total). Uma interpretação possível é que os valores não preenchidos indicam que não houve morte do acidentado. No entanto, não há garantias de que sempre que o valor não é preenchido indica que não houve óbito.
11. **nm\_municipio**: a cidade em que ocorreu o acidente é um atributo categórico nominal e existem 5.285 valores distintos para o esse atributo. O conjunto de dados possui um total de 855.090 instâncias preenchidas com o valor “Não se aplica” (22,04% do total). Embora estejam preenchidos com o valor “Não se aplica”, trata-se de instâncias que não foram corretamente preenchidas.
12. **nome\_uf**: a Unidade Federativa é o estado em que ocorreu o acidente, atributo do tipo categórico nominal. Assim como para o atributo “nm\_municipio”, possui um total de 855.090 instâncias preenchidas com o valor “Não se aplica” (22,04% do total). Embora estejam preenchidos com o valor “Não se aplica”, trata-se de instâncias que não foram corretamente preenchidas.
13. **ds\_natureza\_lesao**: o atributo natureza da lesão descreve o tipo de ferimento ou dano sofrido pelo acidentado. É um atributo categórico nominal e a quantidade de valores ausentes é 854.141 instâncias (22,01%).
14. **ds\_cbo**: a Classificação Brasileira de Ocupações (CBO) descreve e ordena as ocupações em uma estrutura, baseada em informações referentes à natureza e ao conteúdo do trabalho realizado. Existem 2254 valores distintos para o campo CBO. O atributo é do tipo categórico nominal e a quantidade de valores ausentes é 863.185 (22,25% do total).
15. **ds\_parte\_corpo\_atingida**: a parte do corpo atingida descreve o órgão, membro ou região do corpo afetada pelo acidente. Trata-se de um atributo categórico nominal, sem valores ausentes. No entanto, 859.343 instâncias apresentam “Não se aplica” como valor para esse atributo.
16. **cd\_tipo\_sexo\_empregado\_cat**: o sexo do acidentado poderia ser do tipo binário (Masculino ou Feminino), se não fosse por 94 instâncias que foram preenchidas com o valor “Não informado”. Portanto, é do tipo categórico nominal.
17. **ds\_tipo\_acidente**: o tipo de acidente informa como o acidente ocorreu, se ocorreu no exercício do trabalho (acidente típico), se foi durante o trajeto para o trabalho ou se trata de uma doença ocupacional. É um atributo do tipo categórico nominal e 856.307 instâncias tiveram o campo preenchido com o valor “Ignorado”.



18. **ds\_tipo\_local\_acidente**: o local em que ocorreu o acidente é um atributo categórico nominal. Os valores possíveis são: área rural, empregadora, empresa prestadora, via pública ou outros. Um total de 859.343 valores ausentes para esse atributo.

Tabela 4 – Base de Dados de acidentes de trabalho

Atributo	Tipo	Nº ausentes	Observação
Indicador de acidente em feriado	Categórico nominal	3.827.313	-
Agente causador	Categórico nominal	861.008	-
Ano do acidente	Numérico intervalar	0	-
CNAE	Categórico nominal	46.038	292.277 valores preenchidos como “Indefinido”
Data do acidente	Numérico intervalar	0	-
Dia da semana	Categórico ordinal	0	-
Emitente	Categórico nominal	852.877	-
Hora do acidente	Numérico intervalar	0	-
Idade	Numérico racional	173	117 valores questionáveis (idade entre 1 e 14 anos)
Indicador de óbito	Binário	852.862	-
Município	Categórico nominal	0	855.090 valores preenchidos como “Não se aplica”
UF	Categórico nominal	0	855.090 valores preenchidos como “Não se aplica”
Natureza da lesão	Categórico nominal	854.141	-
CBO	Categórico nominal	863.185	-
Parte do corpo atingida	Categórico nominal	0	859.343 valores preenchidos como “Não se aplica”
Sexo	Categórico nominal	0	94 valores preenchidos como “Não informado”
Tipo de acidente	Categórico nominal	0	856.307 valores preenchidos como “Ignorado”
Tipo do local do acidente	Categórico nominal	859.343	-

## 4.2 Método proposto

Este trabalho propõe o uso de técnicas de aprendizado não-supervisionado para a extração de conhecimento desses dados, ou seja, as relações entre os dados será estabelecida pelas suas próprias características, não havendo interferência externa. Essa escolha se justifica pelo fato de a CATWEB não possuir rótulos.

Visto que a intenção desta pesquisa é segmentar a base em grupos de instâncias semelhantes, que poderão prover um ponto de partida para conhecer melhor os dados sobre acidentes de trabalho, optou-se pela tarefa de agrupamento de dados para explorar a CATWEB. Espera-se que, a partir dos grupos de acidentes de trabalho encontrados na base de dados, estes possam ser devidamente caracterizados, auxiliando na tomada de decisão e no emprego de medidas que visam a sua prevenção.

A Figura 5 ilustra as etapas do método proposto para a condução deste trabalho. Nas seções seguintes, serão detalhadas cada uma destas etapas e justificadas cada uma das escolhas realizadas ao longo do processo.

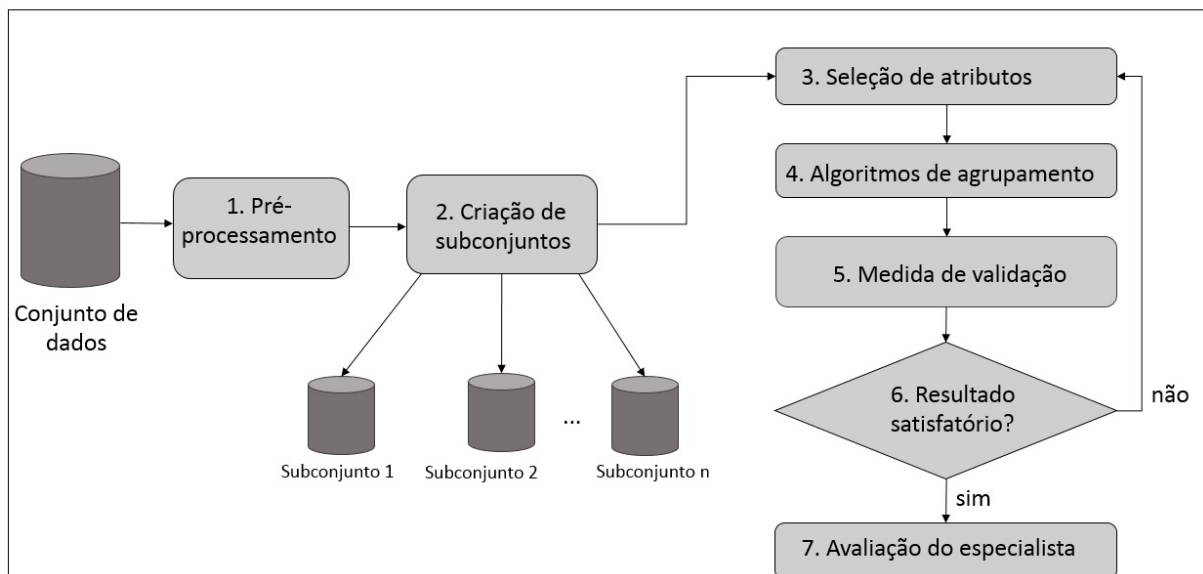


Figura 5 – Etapas do método proposto

### 4.2.1 Etapa 1 - Pré-processamento

Para realizar o pré-processamento, uma versão melhorada da ferramenta de pré-processamento de Silva (2018) foi criada. A ferramenta original foi desenvolvida para trabalhar com a base SISBEN, relativa aos benefícios previdenciários.

A adaptação dessa ferramenta permitiu sua aplicação à base CATWEB. Além disso, novas funções foram implementadas, como: a sumarização de alguns atributos, quebra da base de dados segundo diferentes critérios, conversão de atributos numéricos em categóricos, cálculo do coeficiente de silhueta simplificada.

### 1º passo: Sumarização de atributos

Alguns atributos possuíam uma quantidade muito grande de valores possíveis, fator que compromete a análise baseada em métricas de similaridade aplicadas em espaços euclidianos. Portanto, optou-se por pré-processar esses dados, de forma a sumarizar os seus valores. Os atributos CNAE, Município e Agente causador passaram por esse processo.

O atributo Classificação Nacional de Atividade Econômica (CNAE) é representado por uma estrutura hierárquica, dividida em: seção, divisão, grupo e classe. Na base original, este atributo estava representado pela classe, que consiste na menor granularidade (mais específico), contando com um total de 668 valores distintos. Desta forma, optou-se por aumentar a granularidade, convertendo a classe CNAE para o nível seção, que possui apenas 21 valores diferentes.

O atributo Município teve sua granularidade aumentada, observando-se os critérios de mesorregião estabelecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Mesorregiões são subdivisões dos estados brasileiros em grupos de municípios com semelhanças econômicas e sociais (ESTATÍSTICA-IBGE, 1990). Enquanto a base original contava com 5.285 valores distintos para o atributo Município, após a conversão para Mesorregião, observam-se apenas 165 valores possíveis.

O Agente causador também passou por uma sumarização de seus valores, aumentando sua granularidade. Antes do pré-processamento, existiam 302 valores possíveis para esse atributo e, após a sumarização, esse número reduziu para 21 valores distintos. Ressalta-se que a tabela de referência para essa conversão foi disponibilizada pelos especialistas do Ministério Público do Trabalho.

### 2º passo: Conversão de atributos

O atributo Idade foi tratado segundo duas abordagens distintas: como atributo numérico e como atributo categórico. Para tratá-lo como atributo numérico, nenhum pré-processamento precisou ser aplicado. No entanto, para tratá-lo como categórico, foi necessário sua conversão em faixas etárias.

A seguinte divisão em faixas etárias foi adotada para o pré-processamento:

- ❑  $idade < 18$  anos: Menor de idade
- ❑  $18 \leq idade < 35$ : Jovem Adulto
- ❑  $35 \leq idade < 55$ : Adulto
- ❑  $idade \geq 55$ : Idoso

É importante ressaltar que as instâncias com idade inferior a 16 anos foram removidas da base, como já destacado na seção 4.1.

### 3º passo: Tratamento de valores ausentes

Existem diferentes formas de tratar os valores ausentes, como: eliminar instâncias com valores ausentes, estimar os valores ausentes, ignorá-los ou modificar o algoritmo para ser capaz de lidar com essa questão.

Após avaliar a quantidade de valores ausentes em relação ao total de elementos da base e conversar com os especialistas de domínio, foi considerado razoável eliminar tais instâncias, visto que uma mesma instância possui valores ausentes para vários atributos diferentes, tornando sua presença pouco relevante, ou ainda, causando um impacto negativo para os algoritmos de agrupamento.

Destaca-se que as instâncias que possuíam o atributo Indicador de acidente em feriado com valores ausentes, não foram removidas da base de dados. No entanto, esse atributo não foi utilizado para realizar o agrupamento, pois, quando esse atributo possui valor ausente, não se sabe ao certo se o acidente não ocorreu em feriado ou se atributo não foi devidamente preenchido.

Ao realizar a eliminação das instâncias que possuíam valores ausentes, preservou-se um pouco mais de 80% da base original. Foram mantidas um total de 3.111.563 instâncias.

#### 4.2.2 Etapa 2 - Criação de subconjuntos

Por se tratar de um grande volume de dados, fez-se necessária a divisão da base de dados em subconjuntos para facilitar a análise dos dados e contribuir para o desempenho dos algoritmos de agrupamento. Para isso, foram adotados os seguintes critérios:

- ❑ O primeiro critério adotado para dividir a base de dados foi o ano do acidente. Desta forma, foram criados 6 subconjuntos de dados, um de cada ano (2012 a 2017).
- ❑ O segundo critério adotado para dividir a base de dados foi a mesorregião, totalizando 165 subconjuntos.

#### 4.2.3 Etapa 3 - Seleção de atributos

Por se tratar de uma base de dados com 18 atributos, os especialistas de domínio foram consultados para compreender melhor cada atributo presente na base de dados. Feito isso, optou-se por remover os atributos desnecessários à tarefa de agrupamento ou que trouxessem informações redundantes.

Os atributos Município e UF foram removidos, pois passaram a ser representados pelo novo atributo Mesorregião. Já os atributos Ano e Mesorregião foram utilizados apenas como critérios para subdivisão do conjunto de dados. Portanto, não foram empregados nos experimentos.

Os atributos Data do acidente, Dia da semana, Hora do acidente, Emitente e Indicador de óbito também foram eliminados, pois entendeu-se que as informações fornecidas por esses atributos não seriam de grande relevância para o objetivo desta pesquisa.

Desta forma, os atributos selecionados para compor os subconjuntos de dados utilizados neste estudo, foram: Agente causador, CNAE, Idade, Natureza da lesão, CBO, Parte do corpo atingida, Sexo, Tipo de acidente e Tipo do local do acidente. Outrossim, no Capítulo 5, será descrito um estudo experimental desenvolvido com o objetivo de refinar ainda mais a seleção dos atributos.

#### 4.2.4 Etapa 4 - Algoritmos de agrupamento

Algumas peculiaridades da base de dados foram responsáveis por orientar a escolha dos algoritmos de agrupamento empregados neste estudo, são elas: os diferentes tipos de atributo (numérico e categórico), a presença de um número elevado de categorias para cada atributo categórico e o seu grande volume de dados.

Este trabalho propõe o uso de dois algoritmos de agrupamentos para serem aplicados na base CATWEB, são eles: o HDBSCAN\* (Seção 2.5) e o Cobweb (Seção 2.6). Em relação aos tipos de atributos, o algoritmo Cobweb é capaz de tratar tanto dados categóricos quanto numéricos. Já o HDBSCAN\*, por ser um algoritmo de agrupamento do tipo relacional, trabalha apenas com a distância/dissimilaridade entre as instâncias e não diretamente com os dados. Para isso, foi necessário uma adaptação na forma de realizar o cálculo de distância para tratar tanto atributos numéricos quanto categóricos (Seção 4.3), e assim permitir sua aplicação na CATWEB.

Tanto o HDBSCAN\* quanto o Cobweb são algoritmos de agrupamento hierárquicos. Como a quantidade de grupos é desconhecida *a priori*, a ideia de visualizar uma hierarquia mostrou-se interessante, pois, dessa forma, seria possível avaliar várias possibilidades, com grupos aninhados, e buscar pela melhor solução.

Quanto à questão do grande volume de dados, o Cobweb é interessante, pois realiza apenas uma varredura nos dados, caracterizando-se como um algoritmo com complexidade de tempo de ordem  $O(n)$ , onde  $n$  representa o número de instâncias da base de dados. No entanto, esse tipo de algoritmo pode ser sensível à ordem de entrada dos dados.

O HDBSCAN\* precisa do cálculo da distância entre cada par de instâncias. Neste trabalho, optou-se por calcular a distância entre cada par de instâncias em tempo de execução do código, ao invés de calcular previamente a respectiva matriz de distâncias. Esse cálculo tem complexidade de tempo da ordem de  $O(n^2)$ . Ainda que executando o algoritmo em uma amostra de dados, o tamanho dessa matriz seria muito grande para ser armazenado na memória. No entanto, realizar o cálculo entre cada par de instâncias em tempo de execução aumenta o tempo de execução do algoritmo.

O HDBSCAN\* tem outra característica interessante, além de hierárquico, ele é baseado em densidade. Dessa forma, o formato dos grupos não é uma limitação ao algoritmo, uma vez que ele busca por regiões densas dentro do conjunto de dados.

Já o Cobweb apresenta como diferencial o fato de ser baseado em probabilidades condicionais, valendo-se de uma medida denominada utilidade de categoria (Seção 2.6).

Essa medida utiliza duas probabilidades, enquanto uma verifica qual a proporção de instâncias de um grupo compartilham o mesmo valor para determinado atributo, a outra mensura quantas instâncias de classes distintas compartilham esse mesmo valor.

As adaptações realizadas no algoritmo de agrupamento HDBSCAN\*, bem como as estratégias para realizar o corte na hierarquia do COBWEB e o método empregado para comparar as hierarquias geradas, serão detalhados na Seção 4.3.

## 4.2.5 Etapa 5 - Medida de validação

A medida de validação adotada para verificar a qualidade do agrupamento obtido foi a Silhueta Simplificada (apresentada na Seção 2.8). A sua implementação foi realizada de forma a lidar com as características próprias da base de dados CATWEB.

O coeficiente de Silhueta Simplificada é calculado conforme Equação 11, em que  $b(x_i)$  corresponde à dissimilaridade do  $i$ -ésimo objeto ao centroide do grupo mais próximo, e  $a(x_i)$  é a dissimilaridade do  $i$ -ésimo objeto ao centroide do próprio grupo.

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max[a(x_i), b(x_i)]} \quad (11)$$

O centroide de cada grupo foi obtido da seguinte forma: calculando-se a média dos valores do grupo, no caso dos atributos numéricos; buscando a moda (valor do atributo que mais se repete dentro do grupo), no caso dos atributos categóricos.

Para calcular a distância dos atributos categóricos nominais, foi utilizado o critério de dissimilaridade mostrado na Equação 12.

$$d = \begin{cases} 1, & m \neq n \\ 0, & m = n \end{cases} \quad (12)$$

Conforme a Equação 12, se o valor  $m$  e o valor  $n$ , para determinado atributo, forem iguais, a distância entre eles é igual a zero. Caso contrário, a distância é igual a um.

Para os atributos numéricos, distância é calculada conforme Equação 13.

$$d = \frac{x_a - y_a}{\max_a - \min_a} \quad (13)$$

onde,  $x_a$  e  $y_a$  são os valores para o atributo  $a$  das instâncias  $x$  e  $y$ , respectivamente. O  $\max_a$  representa o maior valor possível para o atributo  $a$  e o  $\min_a$ , o menor valor possível, essa divisão é feita, para que os valores de distância fiquem dentro do intervalo  $[0, 1]$ .

Um atributo específico da base de dados, o dia da semana, requereu uma maneira diferenciada para o cálculo da distância. Para esse caso em particular, foi utilizada uma lista circular. A distância obtida pela Equação 14 corresponde à menor distância entre dois dias da semana, não importando a ordem dos fatores.

$$d = \frac{\min[|x_{diasemana} - y_{diasemana}|, |y_{diasemana} - x_{diasemana}|]}{3} \quad (14)$$

onde,  $x_{diasemana}$  corresponde ao valor do atributo Dia da semana da 1ª instância, e  $y_{diasemana}$  corresponde ao valor do atributo Dia da semana da 2ª instância.

Por fim, para encontrar a distância total entre duas instâncias, é preciso compor a distância quanto a todos os atributos. Neste estudo, adotou-se a média aritmética para tal composição, conforme Equação 15.

$$d(a, b) = \frac{\sum_{i=1}^n d_i(a, b)}{n} \quad (15)$$

onde:  $n$  é o número de atributos do objeto e  $d_i(a, b)$  é a distância entre as instâncias  $a$  e  $b$  em relação ao atributo  $i$ . É importante ressaltar que, da forma como os cálculos foram feitos, as distâncias entre duas instâncias sempre pertencerá ao intervalo  $[0, 1]$ .

O coeficiente de Silhueta Simplificada foi usado para avaliar a qualidade dos diferentes agrupamentos gerados, variando-se os parâmetros do algoritmo HDBSCAN\*, conforme estudo experimental descrito na Seção 5.2. Com essa informação, foi possível escolher a melhor configuração para o algoritmo. A medida de validação também foi utilizada para apoiar a tomada de decisão quanto ao corte na hierarquia 5.4 e para refinar o processo de seleção de atributos 5.3.

#### 4.2.6 Etapa 6 - Verificação dos resultados

Para apoiar a etapa de verificação de resultados, foi utilizada a ferramenta PowerBI<sup>1</sup>. Essa ferramenta foi desenvolvida pela Microsoft e possibilita análise de dados por meio da criação de gráficos e painéis.

O PowerBI foi usado para explorar os grupos encontrados pelos algoritmos de agrupamento. Os gráficos permitem visualizar questões como: a quantidade de instâncias consideradas *outliers*, a quantidade de instâncias por grupo, homogeneidade dos atributos dentro de um grupo e comparar as características de diferentes grupos.

Na etapa de verificação do HDBSCAN\*, são considerados resultados satisfatórios: coeficiente de Silhueta simplificada com valores próximos a 1, grupos homogêneos (verificados com apoio da ferramenta visual) e baixo percentual de *outlier*. Enquanto a verificação do Cobweb baseia-se apenas no coeficiente de Silhueta Simplificada e na análise visual dos grupos, pois este não é capaz de tratar *outliers*.

No caso de os resultados analisados não terem sido satisfatórios segundo tais critérios, uma nova seleção de atributos será feita e todas as etapas seguintes serão repetidas, até que se obtenha um agrupamento bem avaliado.

#### 4.2.7 Etapa 7 - Avaliação do especialista

Após obter agrupamentos que apresentem uma boa avaliação, segundo os critérios verificados na etapa anterior, os resultados serão interpretados e validados por fontes

<sup>1</sup> <https://powerbi.microsoft.com/pt-br/>

externas, como o Observatório Digital de Saúde e Segurança no Trabalho.

Por fim, ao chegar a um conhecimento consolidado, a informação será repassada aos especialistas do Ministério Público do Trabalho, que poderão fazer uso de tal conhecimento para apoiar o processo de tomada de decisões.

## 4.3 Adaptações e estratégias empregadas para uso dos algoritmos de agrupamento

### 4.3.1 HDBSCAN\*

O HDBSCAN\* é um algoritmo de agrupamento do tipo relacional, ou seja, ele não trabalha diretamente com os dados, mas, sim, com a distância entre os elementos do conjunto de dados. Visto que o algoritmo original tratava apenas valores numéricos, foi necessário fazer uma adaptação dos cálculos de distância ou dissimilaridade utilizados, para que o algoritmo fosse capaz de processar a base de dados adotada neste estudo.

O cálculo de distância/dissimilaridade adotado foi o mesmo implementado para a Silhueta Simplificada na Seção 4.2.5. No caso de atributos categóricos nominais, recorreu-se ao critério de dissimilaridade mostrado na Equação 12, para os atributos numéricos, a distância foi calculada conforme Equação 13 e, no caso específico do dia da semana, empregou-se a Equação 14. Por fim, para compor a distância total entre duas instâncias, adotou-se a média aritmética, conforme Equação 15.

### 4.3.2 Cobweb

Diferentemente do HDBSCAN\*, o Cobweb não precisou de adaptações diretamente em seu código, pois ele já é capaz de tratar tanto dados numéricos quanto categóricos. No entanto, em decorrência do fato de o algoritmo fazer apenas uma leitura dos dados, sabe-se que ele pode ser sensível à ordem de entrada dos dados.

Na Seção 6.3, será detalhado um estudo experimental realizado para verificar o comportamento do Cobweb diante da variação da ordem dos dados da amostra. Este estudo recorreu à medida denominada *Hierarchy Agreement Index* - HAI (apresentado na Seção 2.7), que mensura o quanto duas hierarquias são semelhantes entre si. No experimento, a mesma base é apresentada ao algoritmo, variando a ordem de entrada dos dados, gerando assim diferentes hierarquias. Essas hierarquias são comparadas usando o HAI a fim de identificar se o CobWeb produz resultados semelhantes quanto se altera a ordem de apresentação dos dados.

Ademais, o Cobweb não possui o framework FOSC (Seção 2.5.4) para indicar como cortar a hierarquia, a fim de obter os grupos mais significativos, como é o caso do HDBSCAN\*. O trabalho de Bauer (1999) assinala que os conceitos das folhas podem ser inapro-



priados, pois, em alguns casos, contêm apenas uma instância cada, enquanto que os nós que estão próximos da raiz da hierarquia podem apresentar conceitos demasiadamente abstratos, comprometendo a qualidade das informações obtidas.

Assim, os autores do trabalho Bauer (1999) indicam um método para escolha do corte, que eles chamam de “classes apropriadas”. Diz-se que uma classe  $C$ , representada como um nó na árvore, é apropriada se:

- ❑  $C$  não é uma folha;
- ❑ nenhum dos irmãos de  $C$  é um folha;
- ❑ nenhum dos descendentes de  $C$  é um nó apropriado.

Portanto, essa foi a estratégia adotada neste estudo para realizar o corte na hierarquia obtida pelo Cobweb.

## 4.4 Considerações Finais

Este capítulo descreveu a proposta desenvolvida para a execução deste trabalho. Foi detalhado como cada etapa do processo foi aplicada, quais as ferramentas e algoritmos foram empregados, bem como as justificativas para as decisões tomadas ao longo do processo.

Os Capítulos 5 e 6 vão caracterizar cada um dos experimentos realizados e expor os resultados observados.

---

## Experimentos e Análise dos Resultados - HDBSCAN\*

Os experimentos e análises de resultados foram divididos em dois capítulos para facilitar o entendimento. Neste capítulo, serão discutidos os experimentos realizados utilizando o algoritmo HDBSCAN\*. Ao longo das Seções 5.1, 5.2, 5.3 e 5.4 estão detalhadas como foram realizadas as subdivisões dos conjuntos de dados, a forma de parametrização do algoritmo, a escolha dos atributos, o critério para realizar o corte da hierarquia e os impactos observados mediante a variação de tais características. A Seção 5.5 discute os resultados obtidos com a execução do algoritmo HDBSCAN\*. Finalmente, as considerações finais acerca deste capítulo serão discutidas na Seção 5.6.

### 5.1 Divisão da base em subconjuntos de dados

Devido ao grande volume de dados, foi necessário estabelecer critérios para dividir a base em subconjuntos. O primeiro critério adotado foi o ano do acidente. Desta forma, foram criados 6 subconjuntos de dados, um para cada ano (2012 a 2017). Feito isso, foram criadas 6 amostras aleatórias com 40 mil instâncias, uma amostra para cada ano, e os mesmos experimentos foram reproduzidos em cada amostra, a fim de verificar se os agrupamentos gerados variavam ao longo dos anos.

O segundo critério adotado para dividir a base de dados foi o atributo Mesorregião. Após a divisão, foram obtidos 165 subconjuntos. Diante dessa quantidade de subconjuntos, optou-se por analisar dois em especial: o maior subconjunto e o subconjunto referente à mesorregião do Triângulo Mineiro. Foram criadas 2 amostras aleatórias com 40 mil instâncias, uma amostra para cada mesorregião, e os mesmos experimentos foram reproduzidos em cada amostra a fim de estabelecer uma análise comparativa entre os agrupamentos gerados para as duas mesorregiões em estudo.

## 5.2 Parametrização do algoritmo HDBSCAN\*

O HDBSCAN\* possui dois parâmetros de entrada, conforme definido em (CAMPOLLO et al., 2015):

- **minClSize:** o tamanho mínimo dos grupos que se deseja encontrar nas estruturas dos dados analisados. Um grupo ilegítimo caracteriza um subcomponente que tem cardinalidade inferior a MinClSize, então, os objetos desse subcomponente são denominados como *outliers*.
- **minPts:** parâmetro que controla diretamente o tamanho mínimo dos grupos de maneira que os objetos *border* são atribuídos ao mesmo grupo de um dos seus objetos *core*, resultando em grupos que tenham mais de MinPts objetos.

Em Campello et al. (2015) e Campello, Moulavi e Sander (2013), os autores sugerem que seja adotado  $\text{MinClSize} = \text{MinPts}$ , para tornar HDBSCAN\* mais similar à versão original do DBSCAN e simplificar o uso do parâmetro minClSize. Desta forma, o Minpts torna-se um parâmetro único de fator de suavização da estimativa de densidade e, ao mesmo tempo, um parâmetro explícito para controlar o tamanho mínimo dos grupos.

Acatando a sugestão anteriormente mencionada, ainda seria necessário definir qual o melhor valor para tal parâmetro. Para responder a essa questão, utilizou-se uma amostra da base de dados CATBWEB, do ano de 2017, com 40 mil instâncias. Essa amostra foi gerada de forma aleatória, composta pelos atributos definidos na Seção 4.2.3, do Capítulo 4, a saber: Agente causador, CNAE, Idade, Natureza da lesão, CBO, Parte do corpo atingida, Sexo, Tipo de acidente e Local do acidente.

Embora o HDBSCAN\* seja um algoritmo hierárquico, os resultados aqui analisados referem-se a um corte na hierarquia contendo os grupos mais significativos encontrados pelo FOSC. Essa questão será mais bem explorada na Seção 5.4. Para avaliar o impacto da variação dos valores do parâmetro, observou-se o coeficiente de silhueta e o percentual de amostras que foram classificadas como *outlier*. A Tabela 6 ilustra os resultados obtidos.

Tabela 6 – Parametrização HDBSCAN\*

minClSize	minPts	Silhueta	Outlier
2	2	-0,24	62,30%
3	3	0,29	84,65%
<b>4</b>	<b>4</b>	<b>0,52</b>	<b>91,93%</b>
5	5	-0,45	85,65%
6	6	-0,28	87,13%
7	7	-0,39	90,67%

Para os valores de minClSize e MinPts iguais a 2 e 3, o próprio HDBSCAN\* gerou um *warning* sugerindo que aumentasse o valor dos parâmetros. Enquanto, nos valores

minClSize e MinPts superiores a 4, observou-se um decréscimo considerável no valor do coeficiente de Silhueta Simplificada.

Os experimentos para parametrização foram repetidos para outros anos (2012 e 2015), e o comportamento observado foi semelhante ao exposto na Tabela 6, referente ao ano de 2017.

Diante dos resultados observados, optou-se por adotar o  $\text{MinPts} = \text{MinClSize} = 4$  para os próximos experimentos. Ressaltando que esse valor corresponde ao valor padrão sugeridos pelo algoritmo.

No entanto, o coeficiente de *outlier* está muito alto. Acredita-se que isso seja resultado de alguns atributos que estão prejudicando o desempenho do algoritmo na realização do agrupamento. Dessa forma, um estudo da influência de cada um dos atributos no agrupamento dos dados foi realizado.

### 5.3 Estudo dos atributos

Valendo-se dos resultados do experimento de parametrização (Seção 5.2), foi realizado um segundo experimento com o objetivo de analisar a influência dos atributos na qualidade do agrupamento gerado. A pergunta a ser respondida é se um subconjunto do conjunto total de atributos da base poderia trazer melhores resultados para o agrupamento. Para isso, gráficos foram gerados para visualizar os grupos obtidos pelo HDSBSCAN\* para a base CATWEB do ano 2017.

Analisaram-se mais de 20 grupos, de diferentes tamanhos e com diferentes características. Constatou-se que, dentro de cada grupo, os atributos cujos valores não variavam eram: Agente Causador, CNAE, Sexo, Tipo de acidente e Local do Acidente. Portanto, esse foi o ponto de partida para estudar o impacto da variação dos atributos em relação ao agrupamento gerado pelo algoritmo HDBSCAN\*.

O primeiro passo foi executar o algoritmo somente com esses cinco atributos: Agente Causador, CNAE, Sexo, Tipo de acidente e Local do Acidente (conjunto 1). Após, optou-se por adicionar os outros atributos (um a um) e verificar a influência que isso causaria no agrupamento, por meio da medida de validação Silhueta Simplificada.

Na Tabela 7, é possível visualizar a variação do coeficiente de silhueta e do percentual de *outlier* em função dos diferentes conjuntos de atributos utilizados. Ressalta-se que os resultados aqui analisados foram os relativos à solução ideal apontada pelo FOSC.

O primeiro atributo adicionado foi a Idade. Para esse atributo, optou-se por utilizar duas abordagens distintas: convertê-lo em faixas etárias e tratá-lo como categórico (conjunto 2) ou tratar o atributo como numérico (conjunto 3). A partir dessa análise, optou-se por adicionar o atributo idade e adotar a abordagem de faixas etárias. Visto que, embora o coeficiente de silhueta seja igual para os dois casos, o percentual de *outliers*, quando se utilizou a idade como numérico, é maior.

O atributo seguinte a ser adicionado foi o cargo (conjunto 4), cujo resultado não foi satisfatório, visto que o percentual de outlier aumentou consideravelmente, contemplando quase 90% da amostra. Portanto, optou-se por não incluir tal atributo.

O atributo Natureza da Lesão, incluído no conjunto 5, também não foi interessante, uma vez que mais de 90% da amostra foi considerada como outlier.

Por fim, no conjunto 6, verificou-se que ao adicionar o atributo Parte do corpo atingida, o coeficiente de silhueta continua sendo 1 e o percentual de outlier não aumentou muito. Portanto, optou-se por manter esse atributo.

Tabela 7 – Estudo dos atributos

Nº do conjunto	Atributos	Silhueta	Outlier	Observação
1	Agente Causador, CNAE, Sexo, Tipo de acidente e Local do Acidente	1	41,16%	-
2	Agente Causador, CNAE, Sexo, Tipo de acidente, Local do Acidente e idade	1	49,91%	idade como atributo categórico
3	Agente Causador, CNAE, Sexo, Tipo de acidente, Local do Acidente e idade	1	61,04%	idade como atributo numérico
4	Agente Causador, CNAE, Sexo, Tipo de acidente, Local do Acidente, idade e CBO	1	88,43%	-
5	Agente Causador, CNAE, Sexo, Tipo de acidente, Local do Acidente, idade e Natureza da lesão	1	90,47%	-
6	Agente Causador, CNAE, Sexo, Tipo de acidente, Local do Acidente, idade e Parte do corpo atingida	1	51,99%	-

Mediante análise desses resultados, os seguintes atributos foram selecionados para a condução dos próximos experimentos: Agente Causador, CNAE, Sexo, Parte do Corpo Atingida, Idade, Tipo de acidente e Local do Acidente.

É importante ressaltar que existem outras combinações possíveis para realizar o estudo experimental da influência dos atributos. No entanto, a fim de evitar um número alto de combinações, este estudo guiou-se por um caminho inicial que trazia ganhos para o agrupamento e facilitaria a interpretação.

## 5.4 Análise do corte da hierarquia

O HDBSCAN\* utiliza o FOSC para obtenção de grupos significativos a partir da hierarquia gerada. No entanto, o que se observou, ao executar os experimentos das seções 5.2 e 5.3, foi que o corte sugerido pelo FOSC agrupava apenas instâncias com valores iguais para todos os sete atributos empregados na tarefa de agrupamento.

Como já foi explorado, o HDBSCAN\* é um algoritmo de agrupamento hierárquico divisivo, que inicia com todos as instâncias em um único grupo e, ao longo da hierarquia, um grupo é subdividido em grupos menores até que algum critério de parada seja satisfeito. Considerando que, no nível do corte do FOSC, o HDBSCAN\* só agrupava elementos iguais, optou-se por avaliar os níveis acima do corte sugerido pelo FOSC, a fim de verificar o comportamento do algoritmo.

Nos anos de 2012 a 2016, observou-se que, nos níveis acima do corte sugerido pelo FOSC, não houve divisão de grupos, mas apenas a classificação de algumas instâncias como *outliers*, conforme ilustrado na Figura 6.

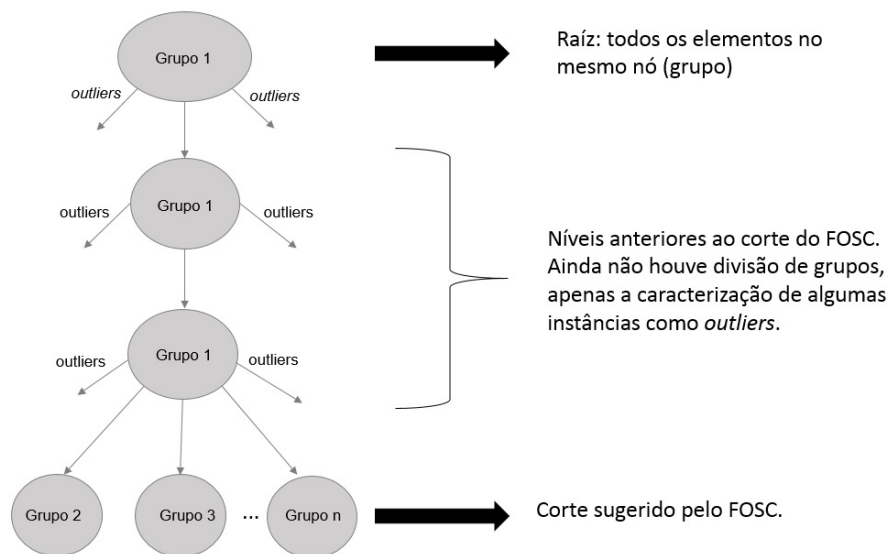


Figura 6 – Hierarquia HDBSCAN\*

Apenas no ano de 2017 o resultado observado foi diferente, conforme se pode observar na Figura 7.

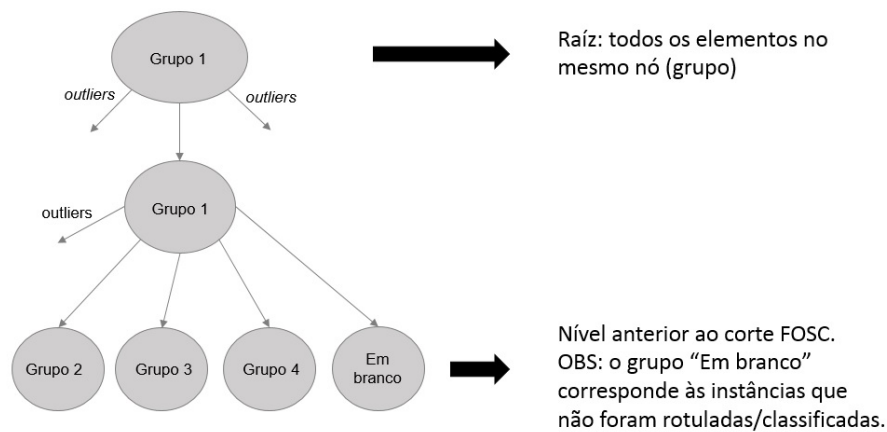


Figura 7 – Hierarquia HDBSCAN\* - Ano 2017

Utilizando o PowerBI, foi possível verificar que o Grupo “em branco” é composto por 7.241 instâncias, essas instâncias não foram alocadas em um grupo válido, tampouco foram consideradas *outliers*. O Grupo 0 refere-se aos *outliers* e possui um total de 901 instâncias. O Grupo 2 consta de 4 instâncias iguais entre si e o Grupo 3 consta de 5 instâncias iguais entre si. Por fim, o Grupo 4 é composto por 31.848 instâncias com valores distintos para os atributos.

A Figura 8 ilustra a composição desses grupos em relação ao percentual de instâncias em cada grupo. Destaca-se que os grupos 2 e 3 possuem apenas 4 e 5 instâncias, respectivamente, por isso, sua visualização é dificultada.

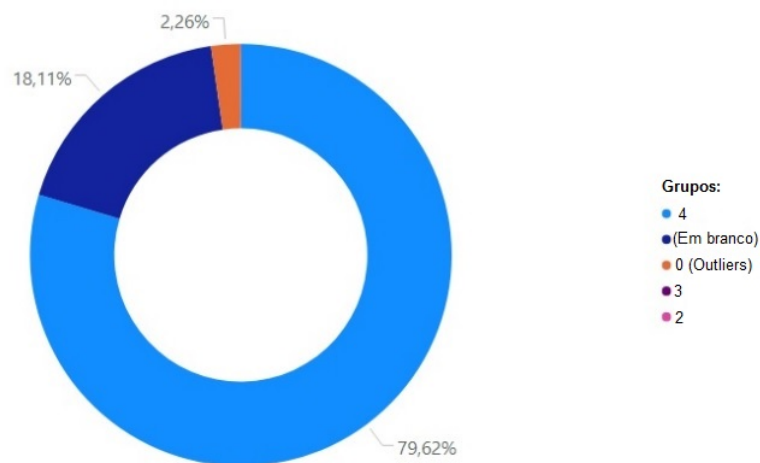


Figura 8 – Grupos 2017 - corte acima do FOSC

Com essa análise, verificou-se que, no corte proposto pelo FOSC, o HDBSCAN\* agrupou apenas instâncias idênticas. No nível acima do FOSC, o algoritmo ainda não havia realizado a divisão de grupos (anos 2012 a 2016, Figura 6) ou 80% das instâncias ainda estavam contidas no mesmo grupo (ano 2017, Figura 7). Dessa forma, optou-se por conti-

nua analisando o corte com os grupos mais significativos, apontado pelo FOSC, e analisar os demais atributos das instâncias (Seções 5.5.1, 5.5.3 e 5.5.4).

Além disso, é importante destacar que os algoritmos baseados em densidade, como o HDBSCAN\*, buscam por regiões de alta densidade, que estejam rodeadas por regiões de baixa densidade (SEMAAN, 2013). Considerando que a amostra gerada aleatoriamente é composta por um grande número de instâncias idênticas, é esperado que o algoritmo considere essas instâncias como regiões de alta densidade. Por isso, observou-se que, no nível indicado pelo FOSC, foram agrupadas apenas as instâncias com valores exatamente iguais. Diante de tal cenário, foram executados experimentos em que instâncias idênticas foram eliminadas da amostra de dados, a fim de verificar o comportamento do algoritmo HDBSCAN\* (Seção 5.5.2).

## 5.5 Resultados do agrupamento usando o HDBSCAN\*

Neste trabalho, objetiva-se analisar e explorar os maiores grupos de acidentes de trabalhos encontrados pelas técnicas de agrupamento. Tais grupos poderiam ser os primeiros a serem analisados pelos especialistas de domínio. Nesta Seção, serão apresentados os resultados dos experimentos realizados ao longo da pesquisa. As Seções 5.5.1 a 5.5.4 detalham como foram conduzidos os experimentos, quais os resultados observados e são tecidas as primeiras conclusões desta pesquisa.

### 5.5.1 Experimento 1:

O Experimento 1 foi executado com o objetivo de caracterizar o maior grupo encontrado pelo HDBSCAN\* na base CATWEB ao dividi-la pelo ano do acidente, buscando responder se o maior grupo se mantém ao longo dos anos e quais são as suas principais características.

Ao dividir a base pelo ano do acidente e considerar apenas os 7 atributos selecionados na Seção 5.3, o agrupamento gerado pelo HDBSCAN\* alcançou o valor máximo para a medida de validação Silhueta Simplificada. Dessa forma, recorreu-se ao PowerBI para visualizar a composição dos grupos formados. Os resultados estão expostos na Tabela 8.

Por meio dessa análise, constatou-se a presença de um grande número de instâncias idênticas, em relação aos atributos selecionados. No entanto, as instâncias na base original não são exatamente iguais, já que existem outros atributos além dos 7 considerados neste experimento.

Como mencionado, a amostra gerada é composta por um grande número de instâncias idênticas, então, o algoritmo considerou essas instâncias como regiões de alta densidade. Por isso, foram agrupadas apenas as instâncias com valores exatamente iguais.



Tabela 8 – Maior grupo de cada ano (HDBSCAN\*)

	2012	2013	2014	2015	2016	2017
<b>Qtde de ins-tâncias</b>	1053	875	935	728	769	704
<b>Sexo</b>	masculino	masculino	masculino	masculino	masculino	masculino
<b>Agente causa-dor</b>	máquinas e equipa-mentos	ferramentas manuais	máquinas e equipa-mentos	máquinas e equipa-mentos	ferramentas manuais	máquinas e equipa-mentos
<b>CNAE</b>	3	3	3	3	3	3
<b>Parte do corpo</b>	dedo	dedo	dedo	dedo	dedo	dedo
<b>Tipo de aci-dente</b>	típico	típico	típico	típico	típico	típico
<b>Local do aci-dente</b>	empregadora	empregadora	empregadora	empregadora	empregadora	empregadora
<b>Faixa etária</b>	adulto	jovem adulto	jovem adulto	jovem adulto	adulto	jovem adulto

Ante tal realidade, optou-se por analisar também os demais atributos dos seis grupos da Tabela 8, referentes ao maior grupo de cada ano, a fim de obter mais informações sobre esses grupos. O seguinte comportamento foi constatado:

- ❑ **Indicador de acidente em feriado:** Em todos os anos, observou-se que mais de 98% das CAT's não ocorreram em feriado;
- ❑ **Dia da semana:** Os acidentes são mais frequentes durante a semana, em comparação ao final de semana, sendo que “Terça-feira” é o dia com mais casos;
- ❑ **Emitente:** Em todos os anos, mais de 90% das CAT's foram registradas pelo empregador;
- ❑ **Indicador de óbito:** analisando o maior grupo de cada ano, constatou-se que não houve nenhum caso de óbito;
- ❑ **Natureza da lesão:** As lesões mais frequentes foram “Corte, laceração, ferida contusa, punctura”, “Fratura” e “Contusão, esmagamento”;
- ❑ **CBO:** a ocupação foi o atributo que apresentou maior diversidade de valores dentro dos grupos. No entanto, para todos os anos, a ocupação com maior registro de CAT's foi “Alimentador de linha de produção”, seguida por “Operador de máquinas fixas em geral” e “Mecânico de manutenção de máquinas”;

Com essa análise, constatou-se que as características dos demais atributos também são muito parecidas, embora não sejam exatamente iguais.

A conclusão do Experimento 1 é que esse tipo de acidente de trabalho acontece com alta frequência. Portanto, verificou-se que trabalhadores do sexo masculino, com idade entre 18 e 34 anos (jovem adulto), que exercem atividades de Pesca e Aquicultura, estão suscetíveis a acidentes durante o exercício de suas atividades e têm o dedo como a parte do corpo lesionada por máquinas e equipamentos ou ferramentas manuais.

Os riscos das atividades de Pesca e Aquicultura (CNAE 3) foram alvo do estudo desenvolvido por Pena e Gomez (2014), no qual os autores destacam a vulnerabilidade desses trabalhadores marcados por condições inseguras, insalubres e sem infraestrutura para proteção à saúde. E, dentre as medidas de prevenção, são citadas a redução da jornada de trabalho excessiva e a disponibilização de equipamentos de proteção individual.

### 5.5.2 Experimento 2:

Considerando apenas os 7 atributos selecionados para o agrupamento, observou-se que a base possui um alto número de instâncias idênticas, fator que pode comprometer o desempenho de algoritmos baseados em densidade, pois as instâncias idênticas são vistas como regiões de alta densidade. Dessa forma, nos experimentos anteriores, o algoritmo HDBSCAN\* foi capaz de agrupar apenas instâncias com valores iguais.

O experimento 2 foi realizado para avaliar o comportamento do HDBSCAN\* após a eliminação das instâncias repetidas da base. A amostra era composta por 53.780 instâncias, das quais 40.646 eram idênticas umas às outras. Após eliminar as instâncias idênticas, foi obtida uma amostra de 13.134 instâncias distintas. E o que se observou foi que o algoritmo não foi capaz de encontrar nenhum agrupamento, todas as instâncias foram classificadas como *outliers*.

Disposto a avaliar e justificar tal comportamento do algoritmo, realizou-se uma análise sobre o funcionamento da medida de distância empregada. A medida de distância proposta neste trabalho produz valores no intervalo entre 0, quando duas instâncias são exatamente iguais, e 1, quando duas instâncias são diferentes em relação a todos os atributos.

Enfatizando que os 7 atributos selecionados neste estudo são categóricos, então, a distância entre cada atributo é sempre 0, para valores iguais, ou 1, para valores diferentes. Ao fazer a média da distância entre os atributos, para obter a distância final entre duas instâncias quaisquer, somente há a possibilidade de 7 valores distintos.

As instâncias da amostra de dados utilizadas neste experimento apresentaram valores sempre iguais a 0,7142 (20,38% da amostra), 0,8571 (39,79% da amostra) ou a 1 (39,83% da amostra). Esses valores indicam que as instâncias apresentam dois, um ou nenhum atributo(s) em comum, respectivamente. Esses altos valores de distância podem ser explicados pelo fato de que todos os atributos selecionados são categóricos e possuem

um grande número de valores distintos. Dessa forma, percebe-se que a amostra contém acidentes de trabalho muito diferentes (dados os atributos selecionados) e o algoritmo não foi capaz de agrupá-las. Além disso, ressalta-se que o algoritmo agrupa procurando por regiões de alta densidade separadas por regiões de baixa densidade. A base em questão pode não apresentar tais características.

A fim de gerar mais valores possíveis de distância entre instâncias, foi realizado um experimento considerando a idade não como atributo categórico, mas como numérico. O novo experimento resultou em valores de distância mais variados e o algoritmo foi capaz de realizar o agrupamento. O HDBSCAN\* segmentou a base em 1954 grupos e alcançou um coeficiente de Silhueta Simplificada igual a 0,57. Todavia, constatou-se que, dentro dos grupos, todas as instâncias eram iguais entre si, exceto em relação ao atributo idade.

Por fim, concluiu-se que a forma como o cálculo da distância entre atributos categóricos está muito rígido, visto que a distância entre dois atributos categóricos diferentes é sempre igual 1. Esse problema poderia ser suavizado, ao estabelecer relações melhores entre as diferentes categorias, resultando em distâncias entre 0 e 1 e não somente em valores necessariamente iguais a 0 ou 1. Em relação ao Agente Causador, por exemplo, “máquinas e equipamentos” é mais semelhante a “ferramentas manuais” e mais dissemelhante a “agente químico”. Dessa forma, seria interessante diminuir a distância entre os mais semelhantes e aumentar a distância entre os mais dissemelhantes. Para isso, faz-se necessária a consulta a um especialista de domínio, que poderia orientar sobre essas similaridades.

### 5.5.3 Experimento 3:

O objetivo do Experimento 3 foi caracterizar os maiores grupos encontrados pelo HDBSCAN\* na base CATWEB ao dividí-la por Mesorregião, buscando comparar duas Mesorregiões distintas, a fim de estabelecer semelhanças e diferenças entre elas.

Ao dividir a base pela mesorregião, foram obtidos 165 subconjuntos. Devido à dificuldade em analisar todos os subconjuntos gerados, optou-se por analisar a mesorregião com o maior número de acidentes e a mesorregião do Triângulo Mineiro, pois é a mesorregião onde foi desenvolvida esta dissertação.

A mesorregião com a maior quantidade de CAT's registradas é a Região Metropolitana de São Paulo. Utilizando os atributos selecionados na Seção 5.3, os parâmetros adotados na Seção 5.2 e uma amostra aleatória de 40 mil instâncias, o HDBSCAN\* encontrou um total de 1.952 grupos.

Analisando os critérios para avaliar a qualidade do agrupamento, observou-se que 49,18% da amostra foi considerada *outlier* e a medida de validação alcançou seu valor máximo. A Tabela 9 sumariza as informações dos 5 maiores grupos encontrados pelo algoritmo HDBSCAN\*.

Tabela 9 – Maiores grupos - Mesorregião 3515 (HDBSCAN\*)

	1º grupo	2º grupo	3º grupo	4º grupo	5º grupo
<b>Qtde de instâncias</b>	754	563	535	272	236
<b>Sexo</b>	masculino	masculino	masculino	feminino	masculino
<b>Agente causador</b>	máquinas e equipamentos	agente biológico	máquinas e equipamentos	agente químico	agente biológico
<b>CNAE</b>	3	8	3	17	8
<b>Parte do corpo</b>	dedo	sistema nervoso	dedo	dedo	sistema nervoso
<b>Tipo de acidente</b>	típico	típico	típico	típico	típico
<b>Local do acidente</b>	empregadora	empregadora	empregadora	empregadora	empregadora
<b>Faixa etária</b>	jovem adulto	adulto	adulto	jovem adulto	jovem adulto

Quando foram analisados os demais atributos dos cinco grupos da Tabela 9, observou-se que:

- ❑ **Indicador de acidente em feriado:** 99,36% dos acidentes não ocorreram em feriado;
- ❑ **Dia da semana:** Os acidentes são mais frequentes durante a semana, em comparação ao final de semana, sendo que “Sexta-feira” é o dia com maior ocorrência;
- ❑ **Emitente:** 99,87% das CAT’s foram registradas pelo empregador;
- ❑ **Indicador de óbito:** não houve nenhum caso de óbito;
- ❑ **Natureza da lesão:** As lesões mais frequentes foram “Corte, laceração, ferida contusa, punctura”, “Doença” e “Fratura”;
- ❑ **CBO:** a ocupação com maior registro de CAT’s foi “Alimentador de linha de produção” (33,26% dos casos), seguida por “Operador de máquinas fixas” (9,36% dos casos);

A mesorregião do Triângulo Mineiro é a 12ª com a maior quantidade de CAT’s registradas. Utilizando os atributos selecionados na Seção 5.3, os parâmetros adotados na Seção 5.2 e uma amostra aleatória de 40 mil instâncias, o HDBSCAN\* encontrou um total de 1.851 grupos.

Destaca-se que 49,42% da amostra foi considerada *outlier* e o valor da Silhueta Simplicada foi igual a um (seu valor máximo). A Tabela 10 sumariza as informações dos 5 maiores grupos encontrados pelo algoritmo HDBSCAN\*.

Tabela 10 – Maiores grupos - Mesorregião 3105 (HDBSCAN\*)

	1º grupo	2º grupo	3º grupo	4º grupo	5º grupo
<b>Qtde de instâncias</b>	792	531	535	272	236
<b>Sexo</b>	masculino	masculino	masculino	feminino	masculino
<b>Agente causador</b>	máquinas e equipamentos	ferramentas manuais	máquinas e equipamentos	ferramentas manuais	agente químico
<b>CNAE</b>	3	3	3	17	3
<b>Parte do corpo</b>	dedo	dedo	dedo	dedo	dedo
<b>Tipo de acidente</b>	típico	típico	típico	típico	típico
<b>Local do acidente</b>	empregadora	empregadora	empregadora	empregadora	empregadora
<b>Faixa etária</b>	jovem adulto	adulto	adulto	jovem adulto	jovem adulto

E, quando foram analisados os demais atributos dos cinco grupos da Tabela 10, observou-se que:

- ❑ **Indicador de acidente em feriado:** 98,61% dos acidentes não ocorreram em feriado;
- ❑ **Dia da semana:** Os acidentes são mais frequentes durante a semana, em comparação ao final de semana, sendo que “Segunda-feira” é o dia com maior ocorrência;
- ❑ **Emitente:** 99,91% das CAT’s foram registradas pelo empregador;
- ❑ **Indicador de óbito:** não houve nenhum caso de óbito;
- ❑ **Natureza da lesão:** As lesões mais frequentes foram “Corte, laceração, ferida contusa, punctura” (47,51%), “Fratura” e “Contusão, esmagamento”;
- ❑ **CBO:** a ocupação com maior registro de CAT’s foi “Magarefe” (13,35% dos casos), seguida por “Alimentador de linha de produção” (11,37% dos casos);

A primeira conclusão desse experimento foi que, assim como no Experimento 1, o algoritmo agrupou apenas instâncias idênticas. Isso se deve ao fato de que a frequência de acidentes com as mesmas características é alta, comprometendo o desempenho do algoritmo baseado em densidade.

A segunda conclusão do Experimento 3 é que as características marcantes observadas quando a base foi dividida por ano (Experimento 1), ainda se mantêm no 1º e no 3º maiores grupos, quando a base é dividida por mesorregião.

Embora apresentem dois grupos idênticos, as duas mesorregiões diferem quanto aos demais grupos. A mesorregião do Triângulo Mineiro apresenta os 5 maiores grupos muito

semelhantes entre si. Todos têm o dedo como parte do corpo atingida, a maioria das vítimas é do sexo masculino e a atividade de Pesca e Aquicultura é a que mais registra acidentes em 4 dos 5 grupos analisados. Enquanto na Região Metropolitana de São Paulo é possível observar grupos mais divergentes. Há maior diversidade quanto a CNAE, Agente causador e parte do corpo atingida. Além disso, as duas mesorregiões diferem quanto ao dia da semana e quanto à ocupação de maior registro de acidentes, mas assemelham-se quanto à natureza da lesão.

Ademais, dois grupos chamam a atenção na Região Metropolitana de São Paulo, o 2º e o 5º maior grupo. Destaca-se que eles se diferenciam entre si apenas quanto a faixa etária. Dessa forma, ao analisar os dois simultaneamente, percebe-se que quase 800 acidentados tiveram o sistema nervoso como parte do corpo atingida. Verificou-se que as características marcantes observadas quando a base foi dividida por ano (Experimento 1), ainda se mantém no 1º e no 3º maiores grupos quando a base é dividida por mesorregião. Esses acidentes acometem, predominantemente, trabalhadores do sexo masculino ao exercer atividades de extração de minerais não-metálicos (CNAE 8). Nesses grupos, não foram registrados óbitos e a natureza da lesão mais frequente é a perda ou diminuição da audição.

Por fim, um grupo composto por vítimas do sexo feminino é verificado nas duas mesorregiões. Esse grupo é composto por trabalhadoras que atuam na fabricação de celulose, papel e produtos de papel (CNAE 17), que tiveram o dedo como parte do corpo atingida. Diferenciando-se apenas quanto ao agente causador.

#### 5.5.4 Experimento 4:

Os grupos encontrados pelo HDBSCAN\*, nos Experimentos 1 e 3, têm o Dedo como a parte do corpo mais frequentemente acometida por acidentes de trabalho. Essa conclusão corrobora a informação obtida no Observatório Digital, onde se verifica que 24% do total de acidentes de trabalho registrados têm o dedo como a parte do corpo atingida, segundo as comunicações de acidentes de trabalho registradas.

O objetivo do Experimento 4 foi analisar os grupos de cada ano, buscando aqueles que não tenham o dedo como a parte do corpo atingida, e assim verificar se as características ainda se mantêm ao longo dos anos.

A Tabela 11 aponta as características do maior grupo de cada ano, cuja parte do corpo atingida não é o dedo.

Percebe-se que, nos anos 2013 e 2016, apesar de a parte do corpo atingida não ser o dedo, as características dos grupos ainda se assemelham muito ao grupo de maior destaque encontrado nos Experimentos 1 e 3. Eles possuem o mesmo CNAE, Agente causador, Faixa etária, Tipo e Local do Acidente. Além disso, a parte do corpo atingida foi a mão, embora se excetuem punhos e dedos.

Nos anos 2012, 2014, 2015 e 2017, pode ser observado o destaque de um grupo com novas características. Os acidentados são do sexo masculino, tiveram o sistema nervoso

Tabela 11 – Grupos com diferentes partes do corpo atingidas

	2012	2013	2014	2015	2016	2017
<b>Qtde de ins-tâncias</b>	112	151	153	215	171	174
<b>Sexo</b>	masculino	masculino	masculino	masculino	masculino	masculino
<b>Agente causa-dor</b>	agente bio-lógico	agente químico	agente bio-lógico	agente bio-lógico	máquinas e equipa-mentos	agente bio-lógico
<b>CNAE</b>	8	3	8	8	3	8
<b>Parte do corpo</b>	sistema nervoso	mão (ex-ceto punho e dedos)	sistema nervoso	sistema nervoso	mão (ex-ceto punho e dedos)	sistema nervoso
<b>Tipo de aci-dente</b>	típico	típico	típico	típico	típico	típico
<b>Local do aci-dente</b>	empregadora	empregadora	empregadora	empregadora	empregadora	empregadora
<b>Faixa etária</b>	jovem adulto	jovem adulto	jovem adulto	adulto	jovem adulto	jovem adulto

como a parte do corpo atingida e o acidente ocorreu ao exercer atividades de extração de minerais não-metálicos (CNAE 8), na própria empregadora. Essas mesmas características foram constatadas no Experimento 2, na Região Metropolitana de São Paulo, nos 2º e 5º maiores grupos da mesorregião. Portanto, acredita-se que esse seja outro grupo que mereça atenção.

A primeira conclusão obtida com o Experimento 4 foi que, mesmo considerando outra parte do corpo atingida pelo acidente de trabalho, as características dos grupos ainda se assemelham ao longo da maioria dos anos. Dessa forma, os trabalhadores do sexo masculino, que trabalham com atividades de Pesca e Aquicultura, devem ficar atentos ao risco de se acidentar e lesionar a mão ou os dedos.

A segunda conclusão extraída é que os trabalhadores do sexo masculino, que exercem atividades de extrações de minerais não metálicos, estão suscetíveis a acidentes causados por um agente biológico, danificando o sistema nervoso e apresentando como principal consequência a perda ou diminuição da audição.

## 5.6 Considerações finais

Este capítulo mostrou como os experimentos foram conduzidos ao longo da pesquisa, descrevendo os desafios e as dificuldades encontradas ao longo do processo.

O primeiro desafio encontrado foi a elevada quantidade de instâncias idênticas, repre-

sentando um grande número de acidentes com características exatamente iguais. Embora essa questão comprometa o desempenho de algoritmos baseados em densidade, como o HDBSCAN\*, o fato de constatar o grande número de acidentes com as mesmas características é uma questão que deve ser repassada ao Ministério Público do Trabalho.

Com o objetivo de superar a dificuldade imposta pelo primeiro desafio, duas abordagens distintas foram seguidas: (i) analisar todos os atributos dos maiores grupos, a fim de obter mais informações sobre eles; (ii) eliminar as instâncias idênticas e executar novamente o algoritmo de agrupamento.

A segunda abordagem proposta, eliminar as instâncias idênticas, conduziu a um segundo desafio: a medida proposta para realizar o cálculo de distância entre atributos categóricos. Verificou-se que a medida não alcançou resultados satisfatórios, pois ela só permite  $n$  valores distintos, quando há  $n$  atributos categóricos, já que as distâncias entre valores categóricos são sempre 0 ou 1. Um estudo mais aprofundado sobre como calcular a distância entre atributos categóricos é uma importante questão a ser trabalhada.

Ademais, quando foi acrescentado um atributo numérico, verificou-se que o algoritmo agrupou instâncias com valores iguais em relação aos atributos categóricos, diferenciando-se apenas em relação ao atributo numérico presente na amostra.

O Capítulo 6 discutirá sobre os experimentos realizados utilizando o Cobweb, estabelecendo relações de semelhanças e divergências entre os resultados dos agrupamentos encontrados com os diferentes algoritmos.



---

## Experimentos e Análise dos Resultados - Cobweb

Neste capítulo, serão discutidos os experimentos realizados utilizando o algoritmo Cobweb. A Seção 6.1 descreve as diretrizes adotadas para a condução dos experimentos realizados e a Seção 6.2 explana sobre a parametrização do algoritmo. Na Seção 6.3, é apresentado um estudo experimental para comparação de hierarquias geradas pelo Cobweb. A Seção 6.4 discute os resultados observados com a execução do algoritmo Cobweb. Por fim, a Seção 6.6 tece as considerações finais deste capítulo.

### 6.1 Diretrizes para condução dos experimentos

Nesta seção, serão detalhados os subconjuntos da base CATWEB, os atributos e os parâmetros utilizados para a realização dos experimentos, além de descrever como foi realizado o corte na hierarquia de grupos obtida com a execução do Cobweb.

- ❑ **Divisão da base em subconjuntos de dados:** os critérios adotados para a divisão da base CATWEB em subconjuntos para realizar os experimentos com algoritmo HDBSCAN\*, foram mantidos para a realização dos experimentos com o algoritmo Cobweb. Então, o primeiro critério utilizado foi o ano do acidente, sendo criados 6 subconjuntos de dados, um para cada ano (2012 a 2017), e o segundo critério adotado foi o atributo Mesorregião, gerando um total de 165 subconjuntos.
- ❑ **Conjunto de atributos:** com a finalidade de possibilitar a comparação dos agrupamentos obtidos com o algoritmo HDBSCAN\*, optou-se por utilizar o mesmo conjunto de atributos definido na Seção 5.3 para realizar a tarefa de agrupamento com o algoritmo Cobweb. Dessa forma, os seguintes atributos foram utilizados para realizar o agrupamento dos dados: Agente Causador, CNAE, Sexo, Parte do CorpoAtingida, Idade, Tipo de acidente e Local do Acidente.

- ❑ **Corte da hierarquia:** Para verificar a qualidade do agrupamento gerado pelo Cobweb e conhecer melhor os grupos mais significativos, foi preciso realizar um corte na hierarquia. Este trabalho recorreu ao método proposto por Bauer (1999), descrito na Seção 4.3.2. O método propõe que um grupo  $C$  é dito apropriado se:  $C$  não é uma folha; nenhum dos irmãos de  $C$  é uma folha; nenhum dos descendentes de  $C$  é um nó apropriado.

Definidas as diretrizes para execução dos experimentos utilizando o Cobweb, realizou-se um estudo para avaliar o impacto da variação da ordem dos dados fornecidos ao algoritmo de agrupamento. Esse estudo será detalhado na próxima seção.

## 6.2 Parametrização do algoritmo Cobweb

O Cobweb possui dois parâmetros de entrada, conforme definido em Gennari, Langley e Fisher (1989):

- ❑ **Cutoff:** controla a amplitude ou fator de ramificação da hierarquia de conceitos produzida. Indica o mínimo valor de utilidade de categoria a ser usado. Um valor baixo para o *cutoff* produz uma árvore com muitos nós-folha, cada um contendo uma única instância, já que nenhuma folha será descartada. Por outro lado, um alto valor de *cutoff* produz uma árvore com nós-folha com muitas instâncias, que representam diferentes grupos.
- ❑ **Acuity:** controla a profundidade de a hierarquia. O *acuity* indica o valor mínimo de desvio padrão para um atributo do grupo.

No entanto, os próprios autores apontam para a fragilidade da variação de tais parâmetros, pois incentivam o ajuste superficial visando alcançar o comportamento desejável. Dessa forma, este trabalho optou por utilizar os valores setados como padrão no *software* Weka (HOLMES; DONKIN; WITTEN, 1994), sendo *acuity* = 1 e *cutoff* = 0.0028.

## 6.3 Comparação entre hierarquias

Algoritmos de agrupamento que realizam uma leitura única do conjunto de dados, como o Cobweb, geralmente, são sensíveis à ordem de entrada dos dados. Esta seção detalhará o estudo experimental realizado com a finalidade de comparar hierarquias produzidas a partir de uma mesma amostra de dados, porém com os dados sendo apresentados em ordens distintas.

Uma amostra aleatória do ano 2017, com 40 mil instâncias, foi selecionada para a realização do estudo. Os dados da amostra foram agrupados pelo algoritmo Cobweb e,

após serem embaralhados, foram novamente submetidos ao agrupamento. Esse procedimento foi repetido 10 vezes, gerando 10 hierarquias diferentes. As hierarquias geradas foram, então, comparadas, duas a duas, usando a medida *Hierarchy Agreement Index* - HAI (Seção 4.3.2) e o resultado obtido está ilustrado na Tabela 12.

Tabela 12 – Comparação de hierarquias usando HAI

	H0	H1	H2	H3	H4	H5	H6	H7	H8	H9
H0	1									
H1	0,96	1								
H2	0,97	0,95	1							
H3	0,99	0,96	0,97	1						
H4	0,97	0,95	0,99	0,97	1					
H5	0,9	0,93	0,92	0,91	0,91	1				
H6	0,96	0,98	0,95	0,96	0,95	0,93	1			
H7	0,95	0,96	0,96	0,95	0,96	0,92	0,96	1		
H8	0,93	0,96	0,94	0,94	0,94	0,95	0,96	0,96	1	
H9	0,97	0,95	0,99	0,97	0,99	0,91	0,94	0,96	0,94	1

Esse estudo foi repetido para mais duas amostras da base CATWEB, uma referente ao ano 2012 e outra referente ao ano 2015. Os resultados observados em todos os testes realizados constataram a medida HAI variando no intervalo  $0,90 \leq HAI \leq 0,99$ .

Levando-se em consideração esses experimentos, concluiu-se que a variação da ordem dos dados não causou um impacto significativo nas hierarquias produzidas pelo algoritmo de agrupamento Cobweb, pois valores de HAI iguais ou maiores de 0,90 indicam hierarquias muito semelhantes.

## 6.4 Resultados do agrupamento usando o Cobweb

Nesta seção, serão apresentados os resultados dos experimentos realizados utilizando o algoritmo Cobweb para o agrupamento dos dados. As Seções 6.4.1 a 6.4.3 detalham como foram conduzidos os experimentos, quais foram os resultados observados e tece as primeiras conclusões desta pesquisa

### 6.4.1 Experimento 5:

O Experimento 5 foi executado com o objetivo de caracterizar o maior grupo encontrado pelo Cobweb\* na base CATWEB ao dividi-la pelo ano do acidente. Mediante tal observação, será possível responder se o maior grupo se mantém ao longo dos anos, quais são as suas principais características e se são semelhantes aos maiores grupos encontrados no Experimento 1 (Seção 5.5.1), que utilizou o algoritmo HDBSCAN\*.

Constatou-se que o coeficiente de Silhueta Simplicada dos agrupamentos analisados, varia entre 0,32 e 0,38 (Tabela 13), ressaltando que, diferentemente do algoritmo HDBS-

CAN\*, o algoritmo Cobweb não trata *outliers*. Dessa forma, todas as instâncias são atribuídas a algum grupo.

Tabela 13 – Coeficiente de Silhueta Simplificada (Cobweb)

	2012	2013	2014	2015	2016	2017
<b>Silhueta</b>	0,35	0,37	0,34	0,32	0,38	0,33
<b>Número de grupos</b>	1298	1129	1298	1235	1184	1234

O Cobweb não agrupou apenas instâncias com valores iguais para os atributos selecionados na Seção 5.3. Os percentuais apresentados na Tabela 14 referem-se aos percentuais do valor mais recorrente de determinado atributo dentro do respectivo grupo.

Tabela 14 – Maior grupo de cada ano (Cobweb\*)

	2012	2013	2014	2015	2016	2017
<b>Qtde de instâncias</b>	1647	1446	1674	1345	1158	1423
<b>Sexo</b>	masculino	masculino	masculino	masculino (99,99%)	masculino	masculino
<b>Agente causador</b>	máquinas e equipamentos	máquinas e equipamentos (70,92%)	máquinas e equipamentos	máquinas e equipamentos	máquinas e equipamentos (91,36%)	máquinas e equipamentos (97,62%)
<b>CNAE</b>	3	3 (87,84%)	3 (70,87%)	3 (89,35%)	3 (98,74%)	3 (94,74%)
<b>Parte do corpo</b>	dedo (79,81%)	dedo (58,36%)	dedo (71,70%)	dedo (90,28%)	dedo (83,43%)	dedo (75,92%)
<b>Tipo de acidente</b>	típico (99,75%)	típico (99,81%)	típico (98,14%)	típico	típico (99,93%)	típico (99,76%)
<b>Local do acidente</b>	empregadora (97,42%)	empregadora	empregadora	empregadora	empregadora	empregadora (91,4%)
<b>Faixa etária</b>	jovem adulto (79,84%)	jovem adulto (90,42%)	jovem adulto (81,24%)	jovem adulto (92,6%)	jovem adulto (82,43%)	jovem adulto (71,91%)

Em relação aos demais atributos, constataram-se as seguintes características:

- ❑ **Indicador de acidente em feriado:** Em todos os anos, observou-se que mais de 95% das CAT's não ocorreram em feriado;
- ❑ **Dia da semana:** Os acidentes são mais frequentes durante a semana, em comparação ao final de semana, sendo que “Segunda-feira” é o dia com mais casos;

- ❑ **Emitente:** Em todos os anos, mais de 95% das CAT's foram registradas pelo empregador;
- ❑ **Indicador de óbito:** analisando o maior grupo de cada ano, constatou-se que não houve nenhum caso de óbito;
- ❑ **Natureza da lesão:** As lesões mais frequentes foram “Corte, laceração, ferida contusa, punctura” e “Contusão, esmagamento”;
- ❑ **CBO:** a ocupação foi o atributo que apresentou maior diversidade de valores dentro dos grupos. No entanto, para todos os anos, a ocupação com maior registro de CAT's foi “Alimentador de linha de produção”, seguida por “Operador de máquinas fixas em geral” e “Mecânico de manutenção de máquinas”;

A conclusão extraída desse experimento é que as características predominantes se repetem no maior grupo de cada ano, são elas:

- ❑ **Sexo:** Masculino
- ❑ **CNAE:** 3 (Pesca e Aquicultura)
- ❑ **Agente causador:** Máquinas e Equipamentos
- ❑ **Parte do corpo atingida:** Dedo
- ❑ **Tipo de acidente:** Típico
- ❑ **Local do acidente:** Empregadora
- ❑ **Faixa etária:** Jovem adulto

### 6.4.2 Experimento 6:

O objetivo do Experimento 6 foi verificar se as características do maior grupo encontrado no Experimento 5, se mantêm quando a base de dados é dividida por Mesorregião e, ainda, comparar as duas Mesorregiões distintas.

Analisando a Região Metropolitana de São Paulo, verificou-se que o Cobweb segmentou a amostra de dados em um total de 1.374 grupos e o valor do coeficiente de Silhueta Simplificada verificado foi igual a 0,35.

A Tabela 15 sumariza as informações dos 5 maiores grupos encontrados pelo algoritmo de agrupamento. Destaca-se que o Cobweb não agrupa apenas instâncias iguais. Portanto, para os atributos que possuem mais de um valor possível dentro do grupo, foi indicado o percentual do valor mais recorrente.

Tabela 15 – Maiores grupos - Mesorregião 3515 (Cobweb)

	1º grupo	2º grupo	3º grupo	4º grupo	5º grupo
<b>Qtde de instâncias</b>	1374	1118	834	662	599
<b>Sexo</b>	masculino	feminino	masculino	masculino	masculino
<b>Agente causador</b>	máquinas e equipamentos	agente biológico	agente biológico (99,64%)	agente químico	máquinas e equipamentos
<b>CNAE</b>	3 (99,85%)	17 (98,34%)	8 (98,92%)	3 (99,85%)	3 (99,83%)
<b>Parte do corpo</b>	dedo	dedo (38,01%)	sistema nervoso	mão (18,43%)	antebraço (10,85%)
<b>Tipo de acidente</b>	típico	típico	típico	típico (99,55%)	típico (99,83%)
<b>Local do acidente</b>	empregadora	empregadora (96,15%)	empregadora	empregadora	empregadora (99,83%)
<b>Faixa etária</b>	jovem adulto (55,02%)	jovem adulto (53,31%)	adulto (68,59%)	jovem adulto (55,14%)	jovem adulto (50,58%)

Constatou-se que as principais características observadas no Experimento 5, ainda se mantêm no 1º e no 5º maiores grupos da Região Metropolitana de São Paulo. O 5º maior diferencia-se apenas quanto à parte do corpo atingida (antebraço). No entanto, ainda é uma parte do corpo localizada nos membros superiores.

Quando foram analisados os demais atributos dos cinco grupos da Tabela 16, observou-se que:

- ❑ **Indicador de acidente em feriado:** 98,95% dos acidentes não ocorreram em feriado;
- ❑ **Dia da semana:** Os acidentes são mais frequentes durante a semana, em comparação ao final de semana, sendo que “Terça-feira” é o dia com maior ocorrência;
- ❑ **Emitente:** 99,69% das CAT’s foram registradas pelo empregador;
- ❑ **Indicador de óbito:** 99,93% dos acidentes registrados não resultaram em óbito;
- ❑ **Natureza da lesão:** As lesões mais frequentes foram “Corte, laceração, ferida contusa, punctura”, “Contusão, esmagamento” e “Doença”;
- ❑ **CBO:** a ocupação com maior registro de CAT’s foi “Alimentador de linha de produção” e “Mecânico de manutenção de máquinas em geral”;

O 4º maior grupo da Região Metropolitana de São Paulo apresentou uma característica ainda não constatada nos demais grupos: foram registrados dois casos de óbitos. As duas vítimas fatais dos acidentes de trabalho aqui verificados eram do sexo masculino. Em um

caso, a parte atingida foi o abdome (inclusive órgãos internos) e, no outro caso, foram afetadas partes múltiplas, ambos acidentes causados por um agente químico.

Buscando por informações relacionadas a acidentes de trabalho que resultaram em óbito, na plataforma do Observatório Digital, verificou-se que as mortes mais recorrentes causadas por agente químico ocorrem quando a parte do corpo atingida é a cabeça, partes múltiplas ou abdome (inclusive órgão internos). Essa informação corrobora a descoberta destacada pelo 4º grupo destacado na Tabela 15.

Acrescenta-se que foram destacados dois grupos na Região Metropolitana de São Paulo no experimento com o HDBSCAN\* (Seção 5.5.2) e esse grupo se repete quando da aplicação do Cobweb. Trata-se do 3º grupo da Tabela 15. A diferença é que o HDBSCAN\* segmentou esse grupo em dois, pela faixa etária, e o Cobweb considerou todos pertencentes ao mesmo grupo.

Analisando a outra mesorregião proposta neste estudo, a mesorregião do Triângulo Mineiro, constatou-se que o Cobweb segmentou a amostra de dados em um total de 1.133 grupos e o valor do coeficiente de Silhueta Simplificada verificado foi igual a 0,32.

A Tabela 16 sumariza as informações dos 5 maiores grupos encontrados pelo algoritmo de agrupamento. O percentual destacado em alguns valores de atributo, corresponde ao percentual do valor mais recorrente para o atributo dentro do respectivo grupo.

Tabela 16 – Maiores grupos - Mesorregião 3105 (Cobweb)

	1º grupo	2º grupo	3º grupo	4º grupo	5º grupo
<b>Qtde de instâncias</b>	1548	1215	845	805	744
<b>Sexo</b>	masculino	masculino	masculino	masculino	masculino
<b>Agente causador</b>	ferramentas manuais	máquinas e equipamentos	agente biológico	máquinas e equipamentos	agente químico (54,17%)
<b>CNAE</b>	3 (98,77%)	3	3	3	10
<b>Parte do corpo</b>	dedo (49,01%)	dedo	dedo (19,17%)	antebraço (12,3%)	dedo
<b>Tipo de acidente</b>	típico (99,87%)	típico (99,84%)	típico (99,64%)	típico (99,88%)	típico (99,33%)
<b>Local do acidente</b>	empregadora (99,74%)	empregadora (98,93%)	empregadora	empregadora	empregadora
<b>Faixa etária</b>	jovem adulto (71,19%)	jovem adulto (65,76%)	jovem adulto (66,15%)	jovem adulto (63,6%)	jovem adulto (66,4%)

Conforme Tabela 16, observa-se que as características destacadas no Experimento 5 ainda se mantêm 1º, 2º e 4º maiores grupos da mesorregião do Triângulo Mineiro quando o Cobweb é utilizado para o agrupamento. O 1º grupo diferencia-se apenas quanto ao Agente causador (ferramentar manuais), e 4º maior grupo, diferencia-se quanto à parte do corpo atingida (antebraço), mas as demais características são as mesmas.

Quando foram analisados os demais atributos dos cinco grupos da Tabela 16, observou-se que:

- ❑ **Indicador de acidente em feriado:** 98,58% dos acidentes não ocorreram em feriado;
- ❑ **Dia da semana:** Os acidentes são mais frequentes durante a semana, em comparação ao final de semana, sendo que “Segunda-feira” é o dia com maior ocorrência;
- ❑ **Emitente:** 99,79% das CAT’s foram registradas pelo empregador;
- ❑ **Indicador de óbito:** 100% dos acidentes registrados não resultaram em óbito;
- ❑ **Natureza da lesão:** As lesões mais frequentes foram “Corte, laceração, ferida contusa, punctura”, “Fratura” e “Contusão, esmagamento”;
- ❑ **CBO:** a ocupação com maior registro de CAT’s foi “Alimentador de linha de produção”, seguida por “Magarefe” e “Mecânico de manutenção de máquinas em geral”;

A primeira conclusão obtida com o Experimento 6 é que as características dos grupos que se destacaram no Experimento 5, ainda são marcantes nos grupos encontrados pelo Cobweb nas duas mesorregiões analisadas.

A segunda conclusão é que, embora apresentem dois grupos idênticos, as duas mesorregiões diferem quanto aos demais grupos. O dia da semana com o maior registro de CAT’s é diferente nas duas mesorregiões. Além disso, na Região Metropolitana de São Paulo foram registrados óbitos. Enquanto no Triângulo Mineiro, nenhum caso de óbito foi registrado nos 5 maiores grupos analisados.

A última conclusão desse experimento diz respeito a um grupo, predominantemente, feminino. O agrupamento realizado com Cobweb encontrou um grupo, predominantemente, feminino apenas na Região Metropolitana de São Paulo. Esse grupo é composto por trabalhadoras que atuam na fabricação de celulose, papel e produtos de papel (CNAE 17), que tiveram o dedo como parte do corpo atingida e um agente biológico como causador do acidente.

### 6.4.3 Experimento 7:

O objetivo do Experimento 7 foi analisar os grupos de cada ano, buscando aqueles que não tenham o dedo como a parte do corpo atingida, e assim verificar se as características ainda se mantêm ao longo dos anos.

A Tabela 17 aponta as características do maior grupo de cada ano, cuja parte do corpo atingida não é o dedo.

Apenas em 2013 a Pesca e Aquicultura não se destaca como CNAE predominante. Nos anos 2012, 2014, 2015 e 2017, observa-se que, mesmo considerando outra parte do corpo



Tabela 17 – Grupos com diferentes partes do corpo atingidas

	2012	2013	2014	2015	2016	2017
<b>Qtde de ins-tâncias</b>	1180	839	1200	969	917	1083
<b>Sexo</b>	masculino	masculino	masculino	masculino	masculino	masculino
<b>Agente causa-dor</b>	ferramentas manuais	agente biológico (99,64%)	ferramentas manuais	agente químico	agente químico	máquinas e equipamentos (99,90%)
<b>CNAE</b>	3 (97,61%)	8 (98,80%)	3 (96,02%)	3	3	3
<b>Parte do corpo</b>	antebraço (18,22%)	sistema nervoso	olho (18,22%)	mão, exceto punho e dedos (16,01%)	pé (14,39%)	mão, exceto punho e dedos (28,16%)
<b>Tipo de acidente</b>	típico (98,89%)	típico	típico (97,69%)	típico (99,49%)	típico (98,25%)	típico (90,90%)
<b>Local do acidente</b>	empregadora (99,91%)	empregadora (99,91%)	empregadora (99,91%)	empregadora	empregadora	empregadora (93,81%)
<b>Faixa etária</b>	jovem adulto (99,57%)	jovem adulto	jovem adulto (99,57%)	jovem adulto (99,48%)	jovem adulto (99,67%)	jovem adulto (91,80%)

atingida pelo acidente de trabalho, as características dos grupos ainda se assemelham ao longo da maioria dos anos. Dessa forma, os trabalhadores do sexo masculino, que trabalham com atividades de Pesca e Aquicultura, devem ficar atentos ao risco de se acidentar e lesionar os membros superiores (destacando mão, dedos e antebraço).

O grupo de 2013 destaca que trabalhadores do sexo masculino, que exercem atividades de extrações de minerais não metálicos, estão suscetíveis a acidentes causados por um agente biológico, danificando o sistema nervoso e apresentando como principal consequência a perda ou diminuição da audição.

## 6.5 Comparação de resultados - HDBSCAN\* x Cobweb

Esta seção estabelece uma análise comparativa entre os resultados observados mediante o uso dos dois diferentes algoritmos de agrupamento.

A primeira análise diz respeito à medida de validação. Os experimentos 1 a 3 realizados com o HDBSCAN\* (Seções 5.5.1 a 5.5.3) apresentaram valores máximos de Silhueta Simplificada. Enquanto o algoritmo Cobweb apresentou valores de Silhueta Simplificada entre 0,32 e 0,38 nos experimentos 5 a 7 (Seção 6.4.1 a 6.4.3). No entanto, destaca-se que os agrupamentos produzidos pelo HDBSCAN\* apresentaram valores de Silhueta Sim-

plificada iguais a 1, pelo fato de que os grupos foram compostos apenas por instâncias idênticas.

Em relação ao tamanho dos grupos, constatou-se que os grupos definidos pelo Cobweb são maiores (em número de instâncias) que os grupos gerados pelo HDBSCAN\*, para uma mesma amostra de dados. Destaca-se, ainda, que a quantidade de grupos definidos pelo Cobweb é inferior. Isso pode ser explicado por dois fatores: o Cobweb não agrupou apenas instâncias exatamente iguais e, também, não rotula nenhuma instância como *outlier*, realizando o agrupamento de todas as instâncias da amostra de dados. A Tabela 18 ilustra um comparativo em relação à quantidade e tamanho de grupos encontrados pelos diferentes algoritmos, para as duas mesorregiões analisadas neste trabalho.

Tabela 18 – Comparação de agrupamentos - HDBSCAN\* x Cobweb

	Meso 3515		Meso 3105	
	HDBSCAN*	Cobweb	HDBSCAN*	Cobweb
Silhueta Simplificada	1	0,35	1	0,32
Número de grupos	1952	1124	1851	1133
Maior grupo	754 instâncias	1374 instâncias	792 instâncias	1548 instâncias
Menor grupo	4 instâncias	4 instâncias	4 instâncias	4 instâncias

Finalmente, as características dos maiores grupos destacados nos experimentos realizados com o HDBSCAN\* se mantêm nos grupos destacados nos experimentos que empregaram o Cobweb. Levando-se a conclusão de que tais grupos realmente se destacam na base de dados CATWEB. São eles:

- ❑ Trabalhadores do sexo masculino, com idade entre 18 e 55 anos, que exercem atividades de Pesca e Aquicultura, que sofrem acidentes de trabalho tendo o dedo como a parte do corpo lesionada por máquinas e equipamentos ou ferramentas manuais.
- ❑ Trabalhadores do sexo masculino, que exercem atividades de extrações de minerais não metálicos, que sofrem acidentes causados por um agente biológico, danificando o sistema nervoso e apresentando como principal consequência a perda ou diminuição da audição;
- ❑ Trabalhadoras do sexo feminino, que atuam fabricação de celulose, papel e produtos de papel, sofrem acidentes que têm o dedo como a parte do corpo atingida. Nesses casos, os agentes causadores mais frequentes são agente químico, agente biológico ou ferramentas manuais;

## 6.6 Considerações Finais

No que tange às descobertas do Observatório Digital de Saúde e Segurança do Trabalho, a presente pesquisa avança ao prover possibilidades de relacionar, de modo automático

e simultâneo, os diversos atributos da base de dados CATWEB. Substitui-se, por meio do uso de algoritmos de agrupamento, o fazer manual, sem direcionamento prévio e que combina somente dois atributos por vez, tal como realizado pela mencionada plataforma.

No entanto, após obter o agrupamento dos dados da CATWEB e destacar alguns grupos, viabilizou-se a busca por informações que pudessem corroborar ou refutar os resultados analisados mediante o uso dessa tarefa da Mineração de Dados. Dessa forma, evidenciaram-se as seguintes conclusões:

- ❑ O grupo que mais se destacou em todos os experimentos realizados, possui o dedo como parte do corpo atingida mais afetada. Na plataforma digital verificou-se que realmente a parte do corpo mais frequentemente atingida é o dedo (24% das CAT's registradas). E, ainda, quando o dedo é atingido, os dois principais agentes causadores são as ferramentas manuais ou as máquinas e equipamentos, características também observadas nos grupos de destaque encontrados pelos dois algoritmos de agrupamentos empregados neste estudo.
- ❑ Analisando todos os experimentos realizados ao longo dessa pesquisa, concluiu-se que os grupos de destaque apresentam como vítimas predominantes, trabalhadores do sexo masculino pertencentes à faixa etária denominada “jovem adulto”. Essas informações corroboram o perfil de casos encontrado no Observatório Digital, onde se verificam que, aproximadamente, 70% dos acidentados são do sexo masculino e mais de 50% das vítimas possuem entre 18 e 34 anos.

Ademais, o trabalho de Rodrigues (2019) concluiu algumas questões com o uso da visualização de dados, que também foram constatadas por meio da análise de grupos. Dentre elas, destaca-se a predominância de “Máquinas e equipamentos” como o agente causador da maior parte de acidentes de trabalho e a recorrência dos membros superiores como a parte do corpo atingida.

Este capítulo mostrou como os experimentos foram conduzidos ao longo da pesquisa, descrevendo quais foram as diretrizes adotadas para realizar os experimentos utilizando o algoritmo Cobweb e detalhando o estudo experimental realizado, para comparar as hierarquias geradas, quando se varia a ordem de entrada dos dados. Foram expostos e discutidos os resultados observados em cada experimento e, por fim, foi realizada uma análise comparativa entre os dois algoritmos de agrupamento empregados nesta pesquisa.

O Capítulo 7 destacará as principais conclusões obtidas nesta dissertação, apontando suas principais contribuições e indicando possíveis trabalhos futuros.

---

## Conclusão

Neste capítulo, serão sumarizadas as conclusões desta pesquisa, destacando as principais contribuições e indicando possíveis trabalhos futuros.

Foi proposta a aplicação de algoritmos de agrupamento com o objetivo de buscar e caracterizar grupos nos dados relativos aos acidentes de trabalho, conferir interpretabilidade aos resultados obtidos, a fim de extrair informações relevantes e por meio dos experimentos realizados ao longo da pesquisa, destacaram-se três grupos de acidente que ocorrem com grande frequência:

- ❑ Trabalhadores do sexo masculino, com idade entre 18 e 34 anos, que exercem atividades de Pesca e Aquicultura, têm o dedo como a parte do corpo lesionada por máquinas e equipamentos ou ferramentas manuais
- ❑ Trabalhadores do sexo masculino, que exercem atividades de extrações de minerais não metálicos e são vítimas de acidentes causados por um agente biológico, danificando o sistema nervoso e apresentando como principal consequência a perda ou diminuição da audição;
- ❑ Trabalhadoras do sexo feminino, que atuam fabricação de celulose, papel e produtos de papel, e sofrem acidentes que têm o dedo como a parte do corpo atingida. Nesses casos, os agentes causadores mais frequentes são agente químico, agente biológico ou ferramentas manuais;

Além disso, as seguintes conclusões foram levantadas:

- ❑ Em geral, as características observadas nos grandes grupos de acidentes de trabalho são as mesmas em todos os anos analisados, nos agrupamentos produzidos pelos dois diferentes algoritmos empregados nesta pesquisa;
- ❑ O agrupamento produzido pelo Cobweb resulta em um número menor de grupos e grupos com um maior número de instâncias, quando comparado ao agrupamento resultante do HDBSCAN\*.

- As características observadas nos grandes grupos de acidentes de trabalho, quando a base de dados é subdividida pelo ano do acidente, se repetem nos grandes grupos de acidente, quando a base é subdividida por mesorregião, para as duas mesorregiões analisadas nesta pesquisa;

Diante dos resultados analisados neste estudo, verificou-se a grande quantidade de acidentes de trabalho com as mesmas características, fator que compromete o desempenho de algoritmos baseados em densidade, como o HDBSCAN\*. Entretanto, ainda se considerou válido destacar os grupos encontrados pelo algoritmo, pois representam os acidentes que ocorrem com mais frequência.

A solução proposta para evitar que o algoritmo HDBSCAN\* agrupasse apenas instâncias idênticas, foi eliminar da amostra de dados as instâncias com valores exatamente iguais. No entanto, observou-se que, ao utilizar apenas os 7 atributos categóricos selecionados, o algoritmo não foi capaz de agrupar nenhuma instância, classificando todas as instâncias como *outliers*. Todavia, ao adicionar um atributo numérico à amostra, o algoritmo agrupou instâncias que se diferenciavam apenas em relação ao respectivo atributo numérico.

Concluiu-se, então, que a forma como se realizou o cálculo da distância entre atributos categóricos foi muito rígido e constatou-se a necessidade de um estudo junto a um especialista de domínio, para poder levantar novos critérios de similaridades entre os dados ou buscar novas formas de sumarizar os atributos categóricos presentes na base de dados em questão.

A Hipótese 2 desta dissertação pôde ser parcialmente constatada, pois as características destacadas nos grandes grupos encontrados no agrupamento dos dados da base CATWEB foram validadas quando se buscaram por fontes externas, como o Observatório Digital de Segurança e Saúde no Trabalho. No entanto, é importante recorrer a outras medidas de validação para verificar se os agrupamentos resultantes consistem em estruturas fortes e indicam grupos de alta relevância.

Por fim, o estudo experimental realizado com a finalidade de comparar hierarquias produzidas pelo Cobweb a partir de uma mesma amostra de dados, porém com os dados sendo apresentados em ordens distintas, atestou que o algoritmo não se mostrou sensível à ordem de entrada dos dados.

## 7.1 Principais Contribuições

Este trabalho apresenta como contribuições a disponibilização de um método para condução de experimentos envolvendo agrupamento de dados; uma ferramenta capaz de realizar o pré-processamento dos dados da base CATWEB; uma medida de cálculo de distância adaptada para calcular a distância entre duas instâncias da base CATWEB

(considerando os seus atributos numéricos e categóricos), e a caracterização dos maiores grupos de acidentes de trabalho.

## 7.2 Trabalhos Futuros

Como trabalhos futuros, indica-se:

- ❑ O uso de medidas de validação distintas, com a finalidade de verificar se os agrupamentos encontrados são realmente relevantes. Sabe-se que coeficiente de Silhueta Simplificada não consiste na melhor forma de avaliar o agrupamento produzido por algoritmos baseados em densidade, como o HDBSCAN\*. Portanto, indica-se o uso de uma medida como a *Density-Based Clustering Validation* (MOULAVI et al., 2014), apropriada para esse algoritmo;
- ❑ O desenvolvimento de técnicas para analisar os diferentes grupos gerados de forma automatizada, pois o número de grupos encontrados é muito grande e torna-se inviável realizar a sua análise de forma manual;
- ❑ O emprego do método proposto nesta pesquisa, também na base de dados SISBEN (relativa aos benefícios previdenciários), com a finalidade de relacionar os grupos de acidentes de trabalho com os gastos levantados pelo INSS;
- ❑ O uso de algoritmos de agrupamento que percorram outras vertentes, como agrupamento em *grids*, o método espectral, agrupamento de setores, tempo e espaço ou, ainda, o uso de mapas de Kohonen;
- ❑ O desenvolvimento de novas formas de calcular a similaridade entre duas instâncias da base CATWEB.

## Referências

- AGGARWAL, C. C. **Data classification: algorithms and applications**. [S.l.]: CRC press, 2014. <https://doi.org/10.1201/b17320>.
- ALMEIDA, P. C. A.; BARBOSA-BRANCO, A. Acidentes de trabalho no Brasil: prevalência, duração e despesa previdenciária dos auxílios-doença. **Revista Brasileira de Saúde Ocupacional**, SciELO Brasil, v. 36, n. 124, p. 195–207, 2011. <https://doi.org/10.1590/S0303-76572011000200003>.
- ALONSO, J. B. Strategies and algorithms for clustering large datasets: a review. **Universitat Politècnica de Catalunya**, 2013. Disponível em: <<https://upcommons.upc.edu/handle/2117/23415>>.
- ALVES, V. S. **Um algoritmo evolutivo rápido para agrupamento de dados**. Dissertação (Mestrado) — Universidade Católica de Santos, 2007. Programa de Mestrado em Informática. Disponível em: <<http://biblioteca.unisantos.br:8181/bitstream/tede/608/1/Vinicius%20Alves.pdf>>.
- ANKERST, M. et al. Optics: ordering points to identify the clustering structure. In: ACM. **ACM Sigmod record**. [S.l.], 1999. v. 28, n. 2, p. 49–60. <https://doi.org/10.1145/304181.304187>.
- ARRUDA, G. F. d. **Uma abordagem de redes complexas para agrupamento de dados**. Monografia (Trabalho de Conclusão de Curso) — UNIVERSIDADE DE SÃO PAULO, 2011. Escola de Engenharia de São Carlos. Disponível em: <<http://www.tcc.sc.usp.br/tce/disponiveis/18/180450/tce-27032012-090330/?&lang=br>>.
- BARTOLOMEU, T. A. **Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento**. Tese (Doutorado) — Florianópolis, SC, 2002. Programa de Pós-Graduação em Engenharia de Produção. Disponível em: <<https://repositorio.ufsc.br/xmlui/handle/123456789/83836>>.
- BATISTA, G. E. d. A. P. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, 2003. Instituto de Ciências Matemáticas e de Computação. Disponível em: <<https://teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/pt-br.php>>.
- BAUER, M. From interaction data to plan libraries: A clustering approach. **International Joint Conferences on Artificial Intelligence**, v. 99, p. 962–967, 1999.

BELLMAN, R. **Adaptive control processes: a guided tour**. [S.l.]: Princeton university press, 2015. v. 2045.

BOENTE, A. N. P.; ROSA, J. L. D. A. Utilização de ferramentas de kdd para integração de aprendizagem e tecnologia em busca da gestão estratégica do conhecimento na empresa. **Simpósio de Excelência em Gestão e Tecnologia–SEGET**. Rio de Janeiro, 2007.

BORGES, V. R. P. Comparação entre as técnicas de agrupamento k-means e fuzzy c-means para segmentação de imagens coloridas. **Encontro Anual de Computação**, 2010. Acesso em: 02/05/2020. Disponível em: <[http://www.enacomp.com.br/2010/cd/artigos/completos/enacomp2010\\_42.pdf](http://www.enacomp.com.br/2010/cd/artigos/completos/enacomp2010_42.pdf)>.

BRACHMAN, R. J.; ANAND, T. The process of knowledge discovery in databases. In: **Advances in Knowledge Discovery and Data Mining**. [S.l.: s.n.], 1996. p. 37–57.

BRAGA, L. P. V. B. **Introdução à Mineração de Dados-2a edição: Edição ampliada e revisada**. [S.l.]: Editora E-papers, 2005.

BRASIL. **Anuário Estatístico da Previdência Social**. 2018. Acesso em: 09/06/2019. Disponível em: <<http://sa.previdencia.gov.br/site/2019/04/AEPS-2017-abril.pdf>>.

\_\_\_\_\_. **Boletim Estatístico da Previdência Social**. 2019. Acesso em: 02/05/2020. Disponível em: <[http://sa.previdencia.gov.br/site/2019/12/Beps102019\\_trab\\_Final\\_\\_PORTAL\\_atualizado.pdf](http://sa.previdencia.gov.br/site/2019/12/Beps102019_trab_Final__PORTAL_atualizado.pdf)>.

BRITO, L. L. **A strategy for temporal visual analysis of labor accident data**. Dissertação (Mestrado) — Universidade Federal de Uberlândia, 2019. Programa de Pós-Graduação em Computação. Disponível em: <<https://repositorio.ufu.br/handle/123456789/28278>>.

CAMPELLO, R. J.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: SPRINGER. **Pacific-Asia Conference on Knowledge Discovery and Data Mining**. [S.l.], 2013. p. 160–172.

CAMPELLO, R. J. et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM New York, NY, USA, v. 10, n. 1, p. 1–51, 2015.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na universidade federal de lavras. **Revista de administração pública**, v. 42, n. 3, p. 495–528, 2008.

CIPA. **Comunicacao de Acidente no Trabalho**. 2011. Acesso em: 18/08/2018. Disponível em: <<http://cipa.fmrp.usp.br/Html/CAT.html>>.

CONCLA. **Comissão Nacional de Classificação**. 2019. Acesso em: 19/05/2019. Disponível em: <<https://cnae.ibge.gov.br/?view=estrutura>>.

DUARTE, F. J. F. **Optimização da Combinação de Agrupamentos baseado na Acumulação de Provas pesadas por Índices de Validação e com uso de Amostragem**. Tese (Doutorado) — Universidade de Trás-os-Montes e Alto Douro, 2008.



ERPLAN. **Observatorio Digital compila dados sobre Seguranca do Trabalho no Brasil. Saiba mais!** 2018. Disponível em: <<http://www.erplan.com.br/noticias/observatorio-digital-compila-dados-sobre-seguranca-do-trabalho-no-brasil-saiba-mais/>>.

ESTATÍSTICA-IBGE, I. B. D. G. E. **Divisão regional do Brasil em mesorregiões e microrregiões geográficas**. [S.l.]: Instituto Brasileiro de Geografia e Estatística Rio de Janeiro, 1990.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **The Knowledge Discovery and Data Mining Conferences**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **Artificial Intelligence Magazine**, v. 17, n. 3, p. 37, 1996.

FAYYAD, U. M. et al. Knowledge Discovery and Data Mining: Towards a unifying framework. In: **The Knowledge Discovery and Data Mining Conferences**. [S.l.: s.n.], 1996. v. 96, p. 82–88.

FERREIRA, G. et al. **Paralelização eficiente de um algoritmo de agrupamento hierárquico**. Monografia (Trabalho de Conclusão de Curso), 2005. Departamento de Computação. Disponível em: <[https://ri.ufs.br/bitstream/riufs/12496/2/Diego\\_Michael\\_Almeida\\_Santana.pdf](https://ri.ufs.br/bitstream/riufs/12496/2/Diego_Michael_Almeida_Santana.pdf)>.

FISHER, D. H. Knowledge acquisition via incremental conceptual clustering. **Machine Learning**, Springer, v. 2, n. 2, p. 139–172, 1987. <https://doi.org/10.1007/BF00114265>.

FLOREK K., J. P. J. S. H. Z. S. Sur la liaison et la division des points d'un ensemble fini. **Colloquium Mathematicum**, v. 2, n. 3-4, p. 282–285, 1951. <https://doi.org/10.4064/cm-2-3-4-282-285>. Disponível em: <<http://eudml.org/doc/209969>>.

FRANCO, N. **Ministério Público lança campanha para prevenir doenças e acidentes de trabalho**. 2018. Acesso em: 10/08/2018. Disponível em: <<http://agenciabrasil.ebc.com.br/geral/noticia/2018-04/ministerio-publico-lanca-campanha-para-prevenir-doencas-e-acidentes-de>>.

FREI, F. **Introdução à análise de agrupamentos**. [S.l.]: Editora UNESP, 2006.

FUNDACENTRO. **Brasil registra 17 mil mortes e 4 milhões de acidentes de trabalho**. 2019. Acesso em: 02/05/2020. Disponível em: <<http://www.fundacentro.gov.br/noticias/detalhe-da-noticia/2019/4/acoes-regressivas-gestao-de-riscos-e-impacto-dos-acidentes-de-trabalho-foram-temas-de-debate>>.

GENNARI, J. H.; LANGLEY, P.; FISHER, D. Models of incremental concept formation. **Artificial intelligence**, Elsevier, v. 40, n. 1-3, p. 11–61, 1989. [https://doi.org/10.1016/0004-3702\(89\)90046-5](https://doi.org/10.1016/0004-3702(89)90046-5).

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining**. [S.l.]: Elsevier Brasil, 2015.

GROSS, J. L.; YELLEN, J. **Graph theory and its applications**. [S.l.]: CRC press, 2005.

- GUHA, S.; MISHRA, N. **Clustering data streams**. [S.l.]: Springer, 2016. 169–187 p. [https://doi.org/10.1007/978-3-540-28608-0\\_8](https://doi.org/10.1007/978-3-540-28608-0_8).
- HAND, D. J. Principles of Data Mining. **Drug safety**, Springer, v. 30, n. 7, p. 621–622, 2007. <https://doi.org/10.2165/00002018-200730070-00010>.
- HINNEBURG, A.; KEIM, D. A. A general approach to clustering in large databases with noise. **Knowledge and Information Systems**, Springer, v. 5, n. 4, p. 387–415, 2003. <https://doi.org/10.1007/s10115-003-0086-9>.
- HOLMES, G.; DONKIN, A.; WITTEN, I. H. Weka: A machine learning workbench. In: IEEE. **Proceedings of ANZIIS'94-Australian New Zealand Intelligent Information Systems Conference**. [S.l.], 1994. p. 357–361.
- HORTA, D. **Algoritmos e técnicas de validação em agrupamento de dados multi-representados, agrupamento possibilístico e bi-agrupamento**. Tese (Doutorado) — Universidade de São Paulo, 2013. Instituto de Ciências Matemáticas e de Computação. Disponível em: <<https://www.teses.usp.br/teses/disponiveis/55/55134/tde-14012014-154211/en.php>>.
- IBGE. **Desemprego recua em dezembro, mas taxa média do ano é a maior desde 2012**. 2018. Acesso em: 01/08/2018. Disponível em: <<https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/19759-desemprego-recua-em-dezembro-mas-taxa-media-do-ano-e-a-maior-desde-2012>>.
- INSS. **Comunicado de Acidente de Trabalho - CAT**. 2018. Acesso em: 11/08/2018. Disponível em: <<https://www.inss.gov.br/servicos-do-inss/comunicacao-de-acidente-de-trabalho-cat/>>.
- JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys (CSUR)**, Acm, v. 31, n. 3, p. 264–323, 1999. <https://doi.org/10.1145/331499.331504>.
- KASZNAR, I. K.; GONÇALVES, B. M. L.; BENTO, M. Técnicas de agrupamento clustering. **Revista Científica e Tecnológica**, 2009.
- LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da Faculdade Salesiana Maria Auxiliadora**, v. 4, p. 18–36, 2009.
- LOBO, R. **Abril Verde**. 2018. Acesso em: 10/08/2018. Disponível em: <<https://www.conceitozen.com.br/abril-verde.html>>.
- MACQUEEN, J. et al. Some Methods for Classification and Analysis of Multivariate Observations. In: OAKLAND, CA, USA. **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.], 1967. v. 1, p. 281–297.
- MENDES, J. C. **Agrupamento de dados e suas aplicações**. 2017. Monografia (Bacharel em Ciência da Computação) - Universidade Federal do Maranhão.
- MENDONÇA, L. O. **Abril Verde: mês dedicado a prevenção de acidentes de trabalho e doenças ocupacionais**. 2017. Acesso em: 10/08/2018. Disponível em: <<http://justificando.cartacapital.com.br/2017/04/12/abril-verde-mes-dedicado-prevencao-de-acidentes-de-trabalho-e-doencas-ocupacionais>>.

- MOULAVI, D. et al. Density-based clustering validation. In: SIAM. **Proceedings of the 2014 SIAM international conference on data mining**. [S.l.], 2014. p. 839–847. <https://doi.org/10.1137/1.9781611973440.96>.
- MPT-MS. **Seminário discute prevencao e doencas ocupacionais**. 2018. Acesso em: 09/08/2018. Disponível em: <[http://portal.mpt.mp.br/wps/portal/portal\\_mpt/mpt/sala-imprensa/mpt-noticias/27ac2b52-631e-4bda-9243-4a599b2c797b](http://portal.mpt.mp.br/wps/portal/portal_mpt/mpt/sala-imprensa/mpt-noticias/27ac2b52-631e-4bda-9243-4a599b2c797b)>.
- MPT-PB. **MPT lanca cartilha de prevencao de acidentes em espacos confinados**. 2018. Acesso em: 09/08/2018. Disponível em: <[http://portal.mpt.mp.br/wps/portal/portal\\_mpt/mpt/sala-imprensa/mpt-noticias/fad8ee53-0fb5-4a80-a55a-6fdccb2a7305](http://portal.mpt.mp.br/wps/portal/portal_mpt/mpt/sala-imprensa/mpt-noticias/fad8ee53-0fb5-4a80-a55a-6fdccb2a7305)>.
- NALDI, M. C. **Técnicas de combinação para agrupamento centralizado e distribuído de dados**. Tese (Doutorado) — Universidade de São Paulo, 2011. Instituto de Ciências Matemáticas e de Computação. Disponível em: <<https://teses.usp.br/teses/disponiveis/55/55134/tde-16032011-113154/en.php>>.
- NETO, A. C. A. et al. Efficient computation and visualization of multiple density-based clustering hierarchies. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, p. 1–1, 2019. <https://doi.org/10.1109/TKDE.2019.2962412>.
- OIT, S. de T. D. M. **Observatório Digital de Saúde e Segurança no Trabalho (MPT-OIT)**. 2017. Dados acessados em: 10/03/2018. Disponível online no seguinte endereço <http://observatoriosst.mpt.mp.br>. Disponível em: <<http://observatoriosst.mpt.mp.br/>>.
- PENA, P. G. L.; GOMEZ, C. M. Saúde dos pescadores artesanais e desafios para a vigilância em saúde do trabalhador. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 19, p. 4689–4698, 2014. <https://doi.org/10.1590/1413-812320141912.13162014>.
- PORTO, S. S. S.; JÚNIOR, I. J. d. N. Acidentes de trabalho no hospital anchieta: Uma análise exploratória de suas características a partir da análise de agrupamentos (clusters). In: **Anais do Congresso Brasileiro de Custos-ABC**. [S.l.: s.n.], 2006.
- RAMOS, C. M.; LOBO, F. Descoberta de conhecimentos em base de dados. **DosAlgarves**, n. 12, p. 53–59, 2003.
- RAND, W. M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical association**, Taylor & Francis Group, v. 66, n. 336, p. 846–850, 1971. <https://doi.org/10.1080/01621459.1971.10482356>.
- RODRIGUES, A. **MPT: A cada quatro horas e meia, uma pessoa morre vitima de acidente de trabalho**. 2018. Acesso em: 18/08/2018. Disponível em: <<http://agenciabrasil.ebc.com.br/geral/noticia/2018-03/mpt-cada-quatro-horas-e-meia-uma-pessoa-morre-vitima-de-acidente-no-brasil>>.
- RODRIGUES, M. P. **A strategy for visual structural data analysis of labor accident data**. Dissertação (Mestrado) — Universidade Federal de Uberlândia, 2019. Programa de Pós-Graduação em Computação. Disponível em: <<https://repositorio.ufu.br/handle/123456789/28282>>.

- ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, North-Holland, v. 20, p. 53–65, 1987. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- SANTANA, V. S. et al. Acidentes de trabalho: custos previdenciários e dias de trabalho perdidos. **Revista de Saúde Pública**, SciELO Public Health, v. 40, p. 1004–1012, 2006. <https://doi.org/10.1590/S0034-89102006000700007>.
- SANTOS, J. A. d. **Algoritmos rápidos para estimativas de densidade hierárquicas e suas aplicações em mineração de dados**. Tese (Doutorado) — Universidade de São Paulo, 2018. Instituto de Ciências Matemáticas e de Computação. Disponível em: <<https://www.teses.usp.br/teses/disponiveis/55/55134/tde-25102018-174244/pt-br.php>>.
- SEMAAN, G. S. **Algoritmos para o Problema de Agrupamento Automático**. Tese (Doutorado) — Tese de Doutorado, Instituto de Computação, Universidade Federal Fluminense, 2013.
- SILVA, D. A. **Aplicação de técnicas de pré-processamento e agrupamento na base de dados de benefícios previdenciários do ministério público do trabalho**. Monografia (Trabalho de Conclusão de Curso), 2018. Faculdade de Computação. Disponível em: <<https://repositorio.ufu.br/bitstream/123456789/22118/1/AplicacaoTecnicaPreprocessamento.pdf>>.
- SOUSA, V. et al. Acidentes de trabalho: custos previdenciários e dias de trabalho perdidos. **Revista de Saúde Pública**, SciELO Public Health, v. 40, p. 1004–1012, 2006. <https://doi.org/10.1590/S0034-89102006000700007>.
- TAN, P.-N. et al. **Introduction to data mining**. [S.l.]: Pearson Education India, 2006.
- TORRES, A. R. A. et al. O adoecimento no trabalho: repercussões na vida do trabalhador e de sua família. **SANARE-Revista de Políticas Públicas**, v. 10, n. 1, 2011.
- VENDRAMIN, L.; CAMPELLO, R. J.; HRUSCHKA, E. R. Relative Clustering Validity Criteria: A comparative overview. **Statistical Analysis and Data Mining: the ASA data science journal**, Wiley Online Library, v. 3, n. 4, p. 209–235, 2010. <https://doi.org/10.1002/sam.10080>.
- WU, X. et al. Top 10 algorithms in Data Mining. **Knowledge and Information Systems**, Springer, v. 14, n. 1, p. 1–37, 2008. <https://doi.org/10.1007/s10115-007-0114-2>.