UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
ESPECIALIZAÇÃO EM BANCO DE DADOS

TIAGO SIEMINKOSKI

DATA MINING: *CLUSTERING* APLICADO A BANCO DE DADOS DE ACIDENTES DE TRABALHO

MONOGRAFIA DE ESPECIALIZAÇÃO

PATO BRANCO
2017

TIAGO SIEMINKOSKI

DATA MINING: *CLUSTERING* APLICADO A BANCO DE DADOS DE ACIDENTES DE TRABALHO

Trabalho de Conclusão de Curso, apresentado ao II Curso de Especialização em Banco de Dados, da Universidade Tecnológica Federal do Paraná, campus Pato Branco, como requisito parcial para obtenção do título de Especialista.

Orientador: Prof. Dr. Dalcimar Casanova.

PATO BRANCO
2017

TERMO DE APROVAÇÃO

DATA MINING: CLUSTERING APLICADO A BANCO DE DADOS DE ACIDENTES DE TRABALHO.

por

TIAGO SIEMINKOSKI

Este Trabalho de Conclusão de Curso foi apresentado em 23 fevereiro de 2017 como requisito parcial para a obtenção do título de Especialista em Banco de Dados. O(a) candidato(a) foi arguido(a) pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Dalcimar Casanova
Prof.(a) Orientador(a)

_____
Fábio Favarim
Membro titular

_____
Viviane Dal Molin de Souza
Membro titular

"O Termo de Aprovação assinado encontra-se na Coordenação do Curso"

AGRADECIMENTOS

À minha família, especialmente para minha esposa Dalila Giovana Pagnoncelli Laperuta, a qual foi fundamental para conclusão desta etapa, apoiando, motivando, contribuindo para conclusão dos trabalhos necessários para cada etapa dessa especialização. Também ao meu filho Samuel Pagnoncelli Sieminkoski que superou minha falta de paciência com muito amor. A todos que, diferente dos finais de semanas costumeiros, entenderam minhas ausências nas comemorações familiares.

À equipe de professores, em especial ao professor Dr. Marcelo Teixeira que levou a especialização até o final, motivando e contribuindo com vários atendimentos para sanar dúvidas.

Ao orientador, professor Dr. Dalcimar Casanova, que mesmo quase sem tempo aceitou o desafio da orientação e teve fundamental importância na decisão do tema deste trabalho.

Aos colegas, que contribuíram com seus conhecimentos, em especial ao Fábio Luiz Uberti que foi meu parceiro em muitos trabalhos e também cedeu alguns ensinamentos.

À Deus, que deu providencias para colocar todas essas pessoas em meu caminho e fortaleceu-os com paciência e a mim força. Obrigado.

Pouco conhecimento faz com que as criaturas se sintam orgulhosas. Muito conhecimento, que se sintam humildes. É assim que as espigas sem grãos erguem desdenhosamente a cabeça para o céu, enquanto que as cheias a baixam para a terra, sua mãe.

Leonardo da Vinci

RESUMO

SIEMINKOSKI, Tiago. Data Mining: Clustering aplicado a banco de dados de acidentes de trabalho. 2017. 36 f. Monografia (II Curso de Especialização em Banco de Dados) - Universidade Tecnológica Federal do Paraná. Pato Branco, 2017.

O excesso de dados disponíveis nas empresas, internet, intranet, por si só não são informações, pelo menos não de uma forma clara. Algumas técnicas de mineração de dados colaboram para que esses dados se convertam em informações úteis que possam ajudar empresas, órgãos públicos, gestores em geral no processo de tomada de decisão. Os acidentes de trabalho são um exemplo de banco de dados com muitas informações, que são organizadas segundo indicadores e contém informações importantes sobre acidentes, acidentados e empresas. Extrair essas informações de um conjunto de dados públicos referentes a sete indicadores de acidentes de trabalho do ano de 2014, pode ajudar a entender as ocorrências dos fatos, identificando grupos ou associações de dados com comportamentos similares.

Palavras-chave: Mineração de dados. Clusterização. Banco de dados. Tomada de decisão. Acidentes de trabalho.

ABSTRACT

SIEMINKOSKI, Tiago. Data Mining: Clustering applied to a workplace accidents database. 2017. 36 f. Monography (II Specialization Course in Database) - Federal University of Technology - Parana. Pato Branco, 2017.

The excess data available in business, internet, intranet, by itself are not information, at least not in a clear way. Some data mining techniques collaborate to make this data useful information that can help companies, public agencies, and managers in the decision-making process. Workplace accidents are an example of a database with lots of information, which are organized according to indicators and contain important information about accidents, accidents and companies. Extracting this information from a set of public data on seven indicators of occupational accidents in 2014 may help to understand the occurrence of events by identifying groups or associations of data with similar behavior.

Keywords: Data Mining. Clustering. Database. Decision-making. Workplace accidents.

LISTA DE SIGLAS

AEAT	Anuário Estatístico de Acidentes de Trabalho
CAT	Comunicado de Acidentes de Trabalho
CNAE	Classificação Nacional de Atividade Econômica
DM	*Data Mining*
G_CNAE	Grupamento de CNAE
KDD	*Knowledge Discovery in Database*
MT	Ministério do Trabalho
SGBD	Sistemas Gerenciadores de Banco de Dados
SI	Sistemas Inteligentes
SQL	*Structured Query Language*

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O avanço tecnológico contribui para o registro e a manipulação de dados, gerando bases populosas. Dessas bases são extraídos indicadores totalizados por categoria e ordem temporal, os quais são utilizados para monitorar eventos e embasar ações. Neste cenário, com o crescimento amplo e constante dessas bases, tornou-se necessário compreender o comportamento dos eventos observados e suas relações com indicadores internos e externos. Para tanto, surgiram as técnicas de mineração de dados (*data mining*), que buscam analisar agrupamentos e associações, buscando encontrar relações entre diferentes indicadores.

No Brasil, diversos órgãos monitoram indicadores nacionais relacionados à educação, saúde, economia, dentre diversos outros. Dentre estes indicadores, estão os índices relacionados aos acidentes de trabalho, tabulados e disponibilizados em forma de relatórios e anuários estatísticos (AEAT, 2014). Diversos estudos investigam, ao longo dos anos, esses indicadores, buscando fotografar o cenário e encontrar formas de redução desses acidentes. Quando tratados de forma quantitativa, indicam a linha de crescimento, declínio ou estabilidade de um determinado índice. Entretanto, quando analisados por similaridade, é possível formar *clusters* que indicam grupos que carecem das mesmas ações preventivas. Portanto, encontrar esses perfis pode orientar ações que podem auxiliar na redução dos índices em um cenário, no qual um acidente a menos já pode ter alta significância.

Portanto, esta monografia se ocupa da aplicação de técnicas de mineração de dados, em banco de dados de acidentes de trabalho no Brasil, de 2014, em busca de novas informações.

1.1 CONSIDERAÇÕES INICIAIS

Com o uso constante de sistemas de informação nas últimas décadas, muitos dados são gerados, criando uma necessidade de desenvolver técnicas de manipulação de dados, de forma que as informações não explícitas sejam extraídas e utilizadas para a compreensão do assunto explorado gerando novas informações. Entretanto, é importante conhecer as diferentes técnicas e escolher as mais adequadas de acordo com as características dos dados em análise. Embora diversas pesquisas utilizem-se dessas técnicas, as mesmas carecem de análises mais profundas a partir dos resultados, sendo tão relevantes quanto aplicar a técnica, compreender o que os dados expressam após as compilações.

1.2 OBJETIVOS

Todo acidente de trabalho tem um registro, há dados desses acidentes em vários lugares, porem o Ministério do Trabalho (MT) criou um documento que reúne todas essas informações o Anuário Estatístico de Acidentes de Trabalho (AEAT). Este estudo manipula os dados do AEAT e aplica técnica de mineração de dados, resultando em associações não visíveis nos AEATs disponibilizados pelo MT, podendo contribuir com medidas que venham compreender e evitar alguns acidentes de trabalho.

1.2.1 Objetivo Geral

Aplicar técnica de data mining e encontrar associações em banco de dados com informações públicas de acidentes de trabalho.

1.2.2 Objetivos Específicos

1. Aplicar técnica de mineração dos dados;
2. Descobrir se há similaridade de acidentes entre CNAEs;
3. Relatar vantagens e desvantagens da técnica.

1.3 JUSTIFICATIVA

Todos os dias nos deparamos com notícias sobre acidentes de trabalho, alguns de maiores proporções afetando mais pessoas outros menores sendo até superficiais, mas o que está em questão é a vida de um ser humano, que quando exposto a áreas de risco no seu trabalho pode sofrer um acidente grave, mutilante ou fatal. Segundo o Anuário Estatístico de Acidentes de Trabalho (AEAT) de 2014 ocorreram no Brasil 704.136 acidentes de trabalho, sendo que 559.061 desses acidentes podem ser identificados quanto ao tipo/natureza de sua causa. Ainda quanto aos acidentes de trabalho, 2.783 trabalhadores faleceram em decorrência de um acidente e 13.833 ficaram com incapacidade permanente para o trabalho. O custo dos acidentes de trabalho são de difícil mensuração, pois envolvem custos diretos e indiretos (SPERANDIO, 1998). Porém, Pastore (2011) concluiu que para as empresas os custos com os acidentes de trabalho do ano de 2009 foram de 41 bilhões de reais e que somadas ao custo da União totalizaram 71 bilhões de reais, ainda, considera que esses dados foram subestimados.

Este estudo envolve técnicas computacionais em favor da vida dos indivíduos, especialmente dos trabalhadores, ou seja, a partir dos dados do AEAT de 2014, aplica-se uma

técnica de mineração de dados para encontrar agrupamentos de dados, que possam contribuir com a segurança e saúde dos trabalhadores.

1.4 ESTRUTURA DO TRABALHO

O presente estudo foi dividido em 5 (cinco) capítulos: o primeiro, acima, introduz o tema e apresenta os objetivos geral e específicos; o segundo traz as principais referências do tema com seus principais conceitos; o terceiro busca identificar os materiais e métodos através dos passos metodológicos; o quarto traz os resultados e os compara com estudos similares e o quinto expõe as conclusões e considerações finais.

2 BANCO DE DADOS E SISTEMAS INTELIGENTES

Muitas técnicas computacionais surgiram durante guerras ou disputas pelo poder. Contudo o desenvolvimento de novas práticas e de novas utilizações para essas técnicas estão crescendo e contribuindo para tomadas de decisão em ambientes corporativos e órgãos públicos para, inclusive, evitar calamidades.

Nesse capítulo são apresentados os principais conceitos que envolvem banco de dados e acidentes de trabalho, bem como as técnicas de *data mining* que auxiliam a tomada de decisão.

Banco de dados é um sistema computadorizado de manutenção de registros, que se define como uma entidade na qual é possível armazenar dados de maneira estruturada e redundantes, ou seja, um mesmo dado pode ser utilizado e representado várias vezes. Para manipular esses dados, são utilizadas operações/comandos/instruções de inserção, busca, exclusão e edição de registros em sistema computadorizado, escritos em linguagem SQL (*Structured Query Language*). A função de integrar o software com o banco de dados é realizada pelos SGBDs (Sistemas Gerenciadores de Banco de Dados), que incorpora as funções de definição, recuperação e alteração de dados (HEUSER, 2009).

Nas últimas décadas, os bancos de dados foram aperfeiçoados, e as necessidades de diversas áreas, bem como o grande volume de dados gerados pelos processos operacionais, contribuíram para o surgimento de Sistemas Inteligentes (SIs). Esses sistemas são capazes de simular e emular o processo de decisão do ser humano para desempenhar tarefas ou resolver problemas, e ainda aproveitar associações e inferências para trabalhar com problemas complexos, armazenando e recuperando eficientemente uma grande quantidade de informações. Para auxiliar no processo decisório, são dispostas diversas técnicas e metodologias de SIs: aquisição de conhecimento, aprendizado de máquina, redes neurais, lógica fuzzy, computação evolutiva, agentes e multiagentes, mineração de dados e de textos (REZENDE, 2005).

2.1 DATA MINING (DM)

A Mineração de Dados (do inglês, *Data mining - DM*) é um campo multidisciplinar que envolve aprendizado de máquina, estatística, bancos de dados, inteligência artificial, busca de informação e visualização. É definida como o processo de descobrir automaticamente informações úteis em grandes repositórios de dados. Para Bing Liu (2011) é a descoberta de conhecimento em bases de dados (KDD *Knowledge Discovery in Database*).

Isto é comumente definida como o processo de descobrir padrões úteis ou conhecimento de fontes de dados, por exemplo: bancos de dados, textos, imagens, da Web, entre outros. Os padrões devem ser válidos, potencialmente úteis e compreensíveis.

Técnicas de mineração de dados são implantadas para vasculhar grandes bases de dados, a fim de encontrar novos modelos que poderiam permanecem desconhecidos. Eles também fornecem capacidades para prever o resultado de uma observação futura. Também é parte integrante da descoberta de conhecimento em bases de dados, que é o processo geral de converter os dados brutos em informação útil. Este processo consiste de uma série de transformações e etapas, a partir de pré-processamento de dados para pós-processamento de mineração de dados (TAN et al. 2006). Existe atualmente um grande número de técnicas de DM, as quais são apresentadas parcialmente na Tabela 1. Essa classificação por funcionalidade auxilia na escolha da técnica a ser aplicada em um banco de dados de acordo com o resultado esperado pelo pesquisador.

(Continua)

Quadro 1: Diversas técnicas de data mining classificadas por funcionalidade desejada

Funcionalidade	Sub-funcionalidade	Técnica
Análise prévia	Análise de *outliers*	Ferramentas de consulta e técnicas estatísticas
		Indução por árvores de decisão
	Análise de desvios	Ferramentas de consulta e técnicas estatísticas
		Indução por árvores de decisão
	Visualização	Agregações e gráficos diversos
Descobrimento	Classificação	Indução por árvores de decisão
	Análise de associações	Mineração de regras de associação (*Market basket analysis*)
		Mineração de regras de associação booleanas unidimensionais a partir de bancos de dados transacionais
		Mineração de regras de associação em múltiplos níveis a partir de bancos de dados transacionais
		Mineração de regras de associação multidimensionais a partir de banco de dados transacionais e data *warehouse*
		Da mineração de associação à análise de correlação
		Mineração de associação baseada em restrição
	Agrupamento (*clustering*)	Métodos de particionamento
		Métodos hierárquicos
		Métodos baseados em densidade
		Métodos baseado em grid
		Métodos de *clustering* baseados em modelos - abordagem estatística e redes neurais
		Análise de *outliers*

(Conclusão)

Funcionalidade	Sub-funcionalidade	Técnica
Descobrimento	Descrição do conceito (caracterização e comparação)	Sumarização e generalização dos dados baseado em categorização
		Caracterização analítica - análise da relevância do atributo
	Segmentação	Indução por árvores de decisão
	Sumarização e visualização	Agregações e gráficos diversos
	Análise em dados no formato texto	Análise de dados textual e recuperação de informações
		Mineração de textos - classificação de documentos e associação por palavras-chave
Estimação /Predição	Estimação/predição	Regressão linear
		Regressão múltipla
		Regressão não linear
		Regressão logística
		Regressão de Poisson
		Outros modelos de regressão
Classificação	Classificação	Indução por árvores de decisão
		Classificação *bayesiana*
		Classificação por *backpropagation* - redes neurais artificiais
		Classificação baseada em conceitos da mineração de regras de associação
		Classificação por *backpropagation* - redes neurais
		Análise de vizinhança - *k-nearest neighboard*
		Casos baseados em raciocínio
		Algoritmos genéricos
		Abordagem por conjuntos *fuzzy*

FONTE: CÔRTES et al. (2002).

Nesta pesquisa, um banco de dados sobre acidentes de trabalho será analisado, em busca de descobrir grupos de CNAE (Classificação Nacional de Atividades Econômicas) similares por índices de acidentes de trabalho. Para tanto, será aplicado método hierárquico (método de Ward), com verificação da distância euclidiana (teste de similaridade) dos agrupamentos (*clustering*), expresso por meio de um dendograma.

2.1.1 *Clustering* ou Análise de Agrupamentos

Clusters dividem os dados em grupos significativos e/ou úteis, e também são o ponto inicial de várias outras análises. Ao processo de agrupar em conjuntos objetos com as mesmas similaridades, chama-se *clustering*. Um cluster é uma classe em potencial, e "*cluster analysis*" é o estudo de técnicas para encontrar automaticamente essas classes (TAN et al, 2006). Ou seja, a análise de cluster é particionar uma população heterogênea em subgrupos homogêneos.

Há diferentes métodos de *clustering*: de partição, métodos hierárquicos, métodos com base na densidade, com base em grelha, e baseados em modelos. A formação dos clusters e a estrutura dos dados foram consideradas para a seleção do método a ser aplicado neste estudo, optando-se pelos hierárquicos (CASTRO, 2003).

2.1.1.1 Métodos hierárquicos / Ward

O agrupamento hierárquico se caracteriza pelo estabelecimento de uma hierarquia ou estrutura em forma de árvore, interligando os objetos por suas associações representados por meio de um dendograma. Podem ser do tipo aglomerativo (de baixo para cima, agrupando pequenos grupos para ficarem maiores) ou particional (de cima para baixo). A Figura 1 apresenta um exemplo de dendograma, com detalhamento textual de grupos e classes. O eixo vertical representa a similaridade e o eixo horizontal a distribuição dos indivíduos, as letras A B, C e D representam os *clusters*, neste exemplo o *cluster* A é similar ao *cluster* B, os *clusters* A e B são similares ao C e os *cluster* A, B e C são similares ao *cluster* D. Dentro dos *clusters* a ideia da hierarquia por similaridade é a mesma para os elementos.

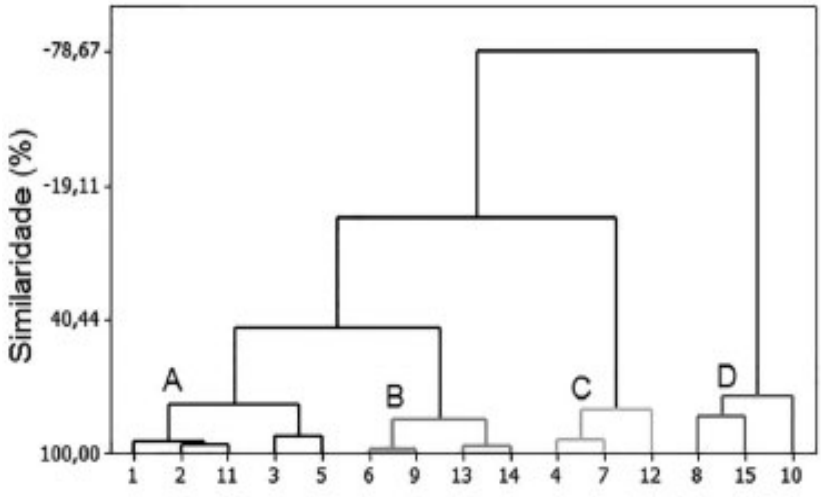

Figura 1: Exemplo de dendograma.
FONTE: CONSTANTINO et al., 2013.

O método de Ward é um procedimento hierárquico no qual, a similaridade para juntar agrupamentos é obtida por meio da soma de quadrados entre os dois agrupamentos somados sobre todas as variáveis. Devido à minimização de variação interna contida nesse método, os agrupamentos resultantes tendem a ter tamanhos aproximadamente iguais. O método Ward, segundo Malhotra (2006) e Kubrusly (2001) é um dos mais utilizados e tem se revelado um dos melhores, métodos hierárquicos de aglomeração. A soma dos quadrados é dada pela fórmula:

$$ESS_k = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

Onde k é o agrupamento em questão, n é o número total de objetos do agrupamento k e xi é o i-ésimo objeto do agrupamento k (Ward, 1963).

A Figura 2 representa as decisões do processo de análise de agrupamentos por meio de diagrama, no qual são listados os métodos mais populares e os critérios que devem ser considerados durante a análise.

No delineamento desta pesquisa, sob orientação do diagrama de decisão (Figura 2), foi utilizado o método hierárquico Ward, e após a análise do dendograma e considerações, os agrupamentos serão interpretados e nomeados, gerando os resultados e discussões sobre os dados minerados.

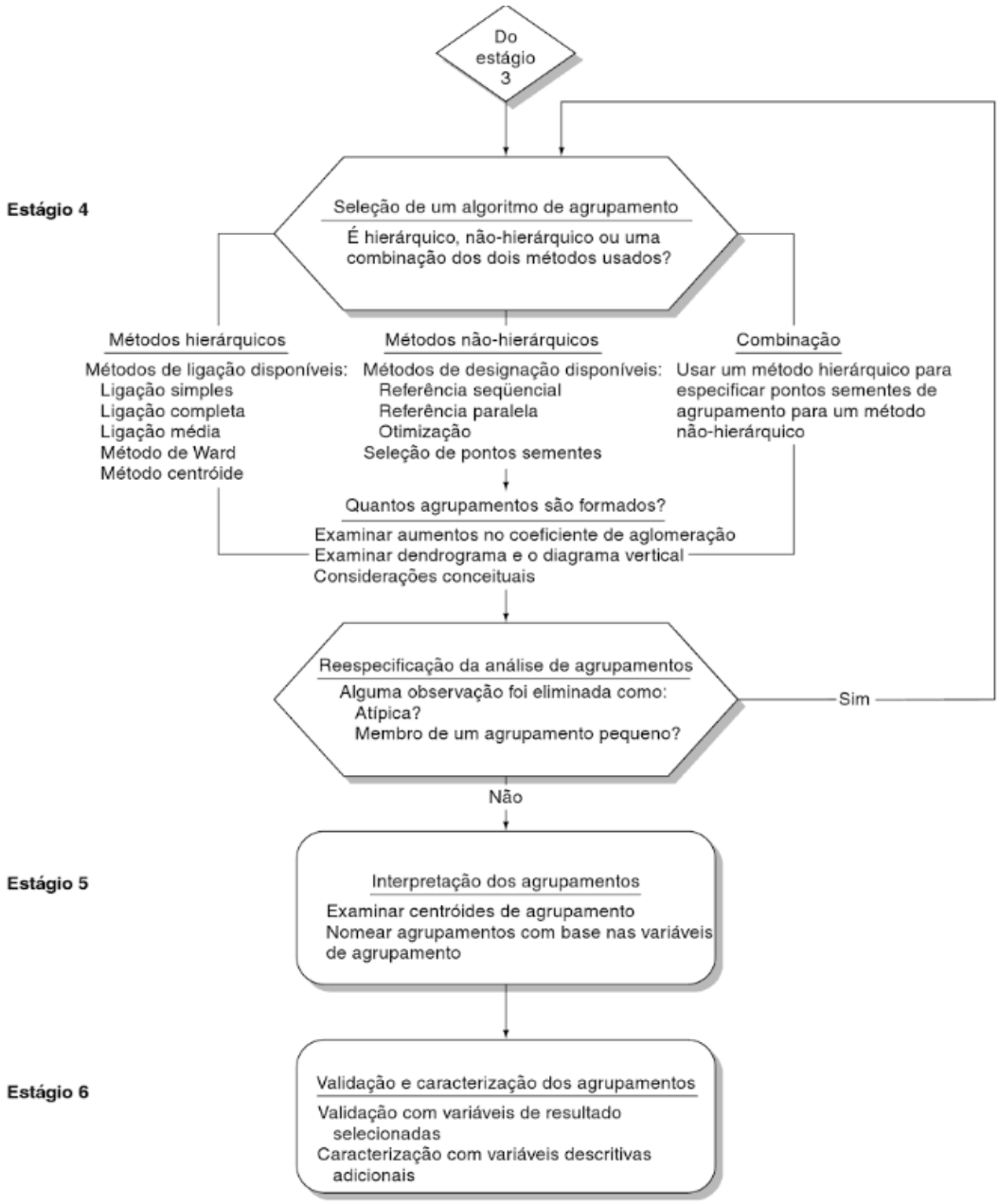

Figura 2: Diagrama de decisão da análise de agrupamentos.
FONTE: HAIR et al., 2009.

2.1.1.2 Distância euclidiana

A distância euclidiana é uma medida que calcula a dissimilaridade entre dois padrões utilizando uma medida de distância definida no espaço de características contínuas (CASTRO, 2003). É a distância mais frequentemente aplicada quando todas as variáveis são quantitativas (SEIDEL et al., 2008), e mede o comprimento em linha reta no desenho entre dois objetos (HAIR, 2009). A Figura 3 apresenta a fórmula para cálculo da distância

euclidiana, representada graficamente pela linha vermelha que, com a aplicação do Teorema de Pitágoras, retorna a distância entre os pontos A e C.

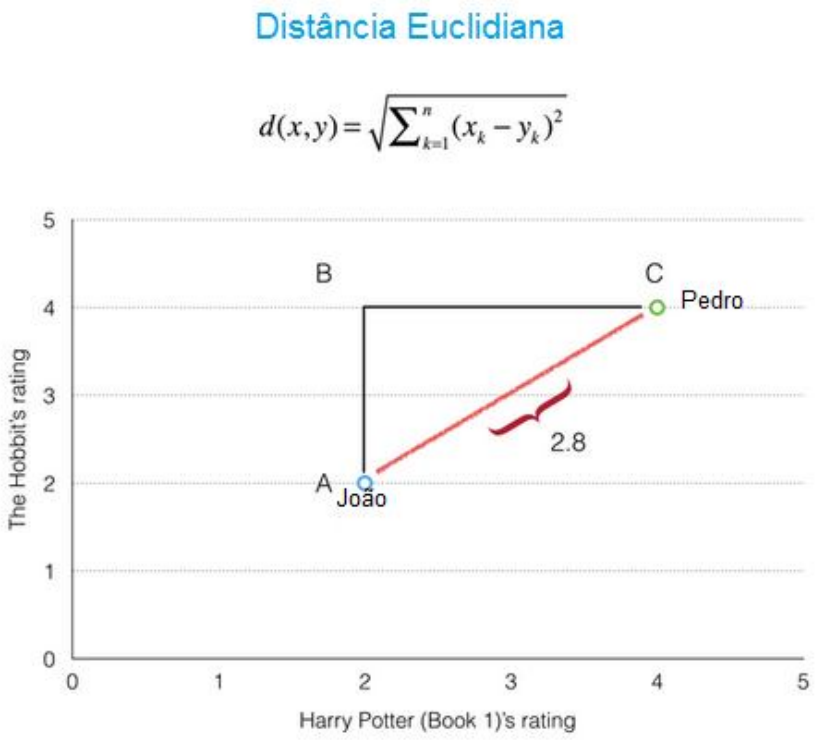

Figura 3: Distância euclidiana, dada pelo Teorema de Pitágoras, em pesquisa de preferências pessoais. FONTE: MOREIRA, 2015.

Essa distância será utilizada para analisar a similaridade entre os grupos de acidentes considerando as variáveis investigadas.

2.2 ACIDENTES DE TRABALHO NO BRASIL

No Brasil, diversos órgãos monitoram indicadores nacionais relacionados à educação, saúde, economia, dentre diversos outros. Dentre estes indicadores, estão os índices relacionados aos acidentes de trabalho, tabulados e disponibilizados em forma de relatórios e anuários estatísticos (AEAT, 2014). Diversos estudos investigam, ao longo dos anos, esses indicadores, buscando fotografar o cenário e encontrar formas de redução desses acidentes.

Nesse contexto o estudo introduz as técnicas supracitadas nos dados do Anuário Estatístico de Acidentes de Trabalho, com intenção de analisar os CNAEs em relação aos indicadores de acidentes de trabalho. Os dados analisados correspondem aos acidentes registrados em 2014 em todo território nacional, com ou sem registro de Comunicado de Acidente de Trabalho – CAT.

A CNAE 2.0 é versão da classificação oficialmente adotada pelo Sistema Estatístico Nacional e pelos órgãos federais gestores de registros administrativos (AEAT, 2014).

O anuário traz informações sobre acidentes de trabalho por gênero, por faixa etária, por tipo de acidente, e ainda por cidade, estado e nação, por atividade econômica (CNAE), e os sete indicadores de acidente de trabalho. Os dados analisados pela técnica de data mining selecionada são os sete indicadores para cada CNAE (662 CNAEs com registro para o ano de 2014). São eles:

1. Incidência (por 1.000 vínculos);
2. Incidência de Doenças Ocupacionais (por 1.000 vínculos);
3. Incidência de Acidentes Típicos (por 1.000 vínculos);
4. Incidência de Incapacidade Temporária (por 1.000 vínculos);
5. Taxa de Mortalidade (por 100.000 vínculos);
6. Taxa de Letalidade (por 1.000 acidentes) e
7. Acidentalidade para a faixa 16 a 34 anos (por 100 acidentes).

A taxa de incidência é um indicador da intensidade com que acontecem os acidentes de trabalho e expressa a relação entre as condições de trabalho e o quantitativo médio de trabalhadores expostos àquelas condições. Seu coeficiente é definido como a razão entre o número de novos acidentes do trabalho a cada ano e a população exposta ao risco de sofrer algum tipo de acidente. Devido à necessidade de publicar os indicadores detalhados por CNAE, decidiu-se pela utilização, no denominador, do número médio de vínculos ao invés do número médio de trabalhadores (AEAT, 2014).

Para a taxa de incidência específica para doenças do trabalho o numerador considera somente os acidentes do trabalho cujo motivo seja doença profissional ou do trabalho, ou seja, aquela produzida ou desencadeada pelo exercício do trabalho, peculiar a determinada atividade e constante de relação existente no Regulamento de Benefícios da Previdência Social (AEAT, 2014).

A taxa de incidência específica para acidentes do trabalho típicos, considera em seu numerador somente os acidentes típicos, ou seja, aqueles decorrentes das características da atividade profissional desempenhada pelo acidentado. Dada a sua natureza é calculada tendo em vista somente os acidentes com CAT registrada, para os quais é possível identificar o motivo do acidente (AEAT, 2014).

Para a taxa de incidência específica para incapacidade temporária, são considerados no numerador, os acidentes do trabalho nos quais os segurados ficaram temporariamente

incapacitados para o exercício de sua capacidade laboral, independentemente da duração do afastamento da atividade. Durante os primeiros 15 dias consecutivos ao do afastamento da atividade, caberá à empresa pagar ao segurado empregado o seu salário integral. Após este período, o segurado deverá ser encaminhado à perícia médica da Previdência Social para requerimento de um auxílio-doença acidentário (AEAT, 2014).

A taxa de mortalidade mede a relação entre o número total de óbitos decorrentes dos acidentes do trabalho verificados no ano e a população exposta ao risco de se acidentar (AEAT, 2014).

Entende-se por letalidade, a maior ou menor possibilidade de o acidente ter como consequência a morte do trabalhador acidentado. A taxa de letalidade é um bom indicador para medir a gravidade do acidente, seu coeficiente é calculado pelo número de óbitos decorrentes dos acidentes do trabalho e o número total de acidentes (AEAT, 2014).

A taxa que mede a acidentalidade, proporcional específica para a faixa etária de 16 a 34 anos tem por objetivo, revelar o risco específico de se acidentar para o subgrupo populacional de trabalhadores na faixa etária de 16 a 34 anos e pode ser expresso como a proporção de acidentes que ocorreram nesta faixa etária em relação ao total de acidentes (AEAT, 2014).

Com base nessas características, esta pesquisa busca descobrir se há similaridade entre atividades econômicas do ponto de vista dos acidentes de trabalho, através dos indicadores, por meio da análise de agrupamentos.

3 MATERIAIS E MÉTODOS

A pesquisa teve como base uma revisão de literatura sobre fundamentos de *Data Mining* e Análise de Agrupamentos. Os dados do AEAT estavam disponibilizados no formato pdf (Figura 4), os quais foram exportados para o Excel, sendo que precisou retirar os cabeçalhos das páginas e o cabeçalho das tabelas e a coluna de CNAEs, também foi necessário atribuir o valor zero (0) no lugar do traço (-) que estava representando ausência de informação resultando em uma tabela constituída por 660 linhas e 7 colunas (660x7) conforme a Figura 5.

59.2 - Indicadores de acidentes do trabalho, segundo a Classificação Nacional de Atividades Econômicas (CNAE), Brasil - 2014

(continua)

CNAE	INDICADORES DE ACIDENTES DO TRABALHO						
	Incidência (por 1.000 vínculos)	Incidência de Doenças Ocupacionais (por 1.000 vínculos)	Incidência de Acidentes Típicos (por 1.000 vínculos)	Incidência de Incapacidade Temporária (por 1.000 vínculos)	Taxa de Mortalidade (por 100.000 vínculos)	Taxa de Letalidade (por 1.000 acidentes)	Acidentalidade para a faixa 16 a 34 anos (por 100 acidentes)
TOTAL	16,06	0,36	9,76	13,66	6,35	3,95	50,54
0111	21,53	0,27	14,39	20,59	46,68	21,68	41,96
0112	26,88	–	22,25	26,88	–	–	53,33
0113	23,46	0,10	18,52	20,33	11,58	4,94	47,34
0114	17,12	0,52	9,86	17,64	–	–	42,42
0115	25,01	0,12	18,83	24,71	45,97	18,38	51,60
0116	13,49	–	8,47	13,10	13,23	9,80	43,14
0119	14,30	0,04	9,60	13,56	7,84	5,48	50,68
0121	16,91	0,29	12,33	16,63	9,55	5,65	56,50
0122	13,19	–	8,33	13,00	–	–	49,62
0131	27,18	0,07	22,94	24,38	14,70	5,41	41,03
0132	30,96	0,07	19,97	30,76	71,54	23,11	47,48
0133	17,07	0,07	10,78	15,86	10,48	6,14	53,84
0134	17,28	0,14	13,52	16,76	16,52	9,56	35,25
0135	14,44	–	7,42	14,24	–	–	47,22
0139	41,47	0,25	36,72	28,38	9,99	2,41	64,46
0141	68,86	0,27	56,92	66,65	73,83	10,72	39,57
0142	21,76	0,27	14,78	20,96	–	–	49,38
0151	20,25	0,16	14,27	19,83	21,78	10,76	40,13
0152	15,46	0,25	11,54	14,73	24,55	15,87	33,33
0153	10,05	1,01	7,04	10,05	100,54	100,00	20,00

Figura 4: Dados do Anuário 2014 formato original pdf (primeiros 20 de 660)

```
21.5255    0.2746   14.3869   20.5920   46.6752   21.6837   41.9643
26.8787         0   22.2496   26.8787         0         0   53.3333
23.4563    0.0993   18.5168   20.3288   11.5833    4.9383   47.3369
17.1191    0.5188    9.8565   17.6379         0         0   42.4242
25.0125    0.1192   18.8317   24.7060   45.9726   18.3799   51.5997
13.4941         0    8.4669   13.0973   13.2296    9.8039   43.1373
14.3046    0.0392    9.6017   13.5599    7.8381    5.4795   50.6849
16.9122    0.2866   12.3258   16.6255    9.5549    5.6497   56.4972
13.1942         0    8.3332   12.9958         0         0   49.6241
27.1783    0.0735   22.9386   24.3845   14.7042    5.4103   41.0280
30.9587    0.0650   19.9670   30.7636   71.5432   23.1092   47.4790
17.0675    0.0699   10.7785   15.8621   10.4816    6.1412   53.8383
17.2779    0.1416   13.5249   16.7586   16.5226    9.5628   35.2459
14.4404         0    7.4208   14.2399         0         0   47.2222
41.4692    0.2498   36.7228   28.3790    9.9926    2.4096   64.4578
68.8637    0.2685   56.9166   66.6488   73.8305   10.7212   39.5712
21.7640    0.2687   14.7780   20.9579         0         0   49.3827
20.2455    0.1633   14.2705   19.8304   21.7767   10.7563   40.1345
15.4649    0.2455   11.5373   14.7284   24.5474   15.8730   33.3333
10.0545    1.0054    7.0381   10.0545  100.5446  100.0000   20.0000
```

Figura 5: Dados limpos e ajustados (primeiros 20 de 660).

Os dados “limpos” foram exportados do Excel para o Matlab, no qual foram criados vetores denominados: X para as colunas dos indicadores pelas linhas (660x7); CNAE2 para a

coluna de CNAEs (660x1); Z que é a aplicação dos métodos Ward e a distância euclidiana em X e H que gera um dendograma de Z por CNAE2, conforme Figura 6.

```
clear; clc; close all;
load AT;
Z = linkage(X,'ward','seuclidean');
figure();
H = dendrogram(Z,0,'labels',CNAE2);
```

Figura 6: Código do matlab para geração do dendograma.

No código apresentado, foram utilizados os comandos de limpeza: *clear* (limpa a *workspace)*, *clc* (limpa a janela de comando) e *close all* (fecha as *Figure Windows*[1] abertas). O comando *load AT* carrega um grupo de vetores, que traz os dados sobre os indicadores de acidente de trabalho (X, Z, H, CNAE2). Já a expressão Z = linkage (X,method,metric) executa o *clustering* usando a métrica de medida de distância para calcular distâncias entre as linhas de X. O comando figure ( ) abre uma nova *Figure Window*, onde o próximo gráfico plotado será mostrado. Por fim, a instrução H = dendograma (___) gera um gráfico de dendograma e retorna um vetor. A Tabela 1 apresenta as ferramentas e tecnologias utilizadas:

Tabela 1: Ferramentas e tecnologias utilizadas.

NOME	VERSÃO	DESCRICÃO DE USO
EXCEL	Microsoft Office Professional Plus 2013	Organização, limpeza e classificação de indicadores por meio de planilhas eletrônicas
MATLAB	R2015a	Clusterização, por meio da aplicação do método Ward, e geração do dendograma.

[1] A *Figure Window* é a janela que exibe gráficos no Matlab.

4 RESULTADOS E DISCUSSÃO

Os sete indicadores (apresentados no capítulo 2.2), referente ao ano de 2014 foram analisados por meio do método hierárquico Ward, que une os elementos por meio da soma de quadrados da similaridade, resultando em um dendograma com a similaridade dos CNAEs por meio da distância euclidiana, conforme Figura 7:

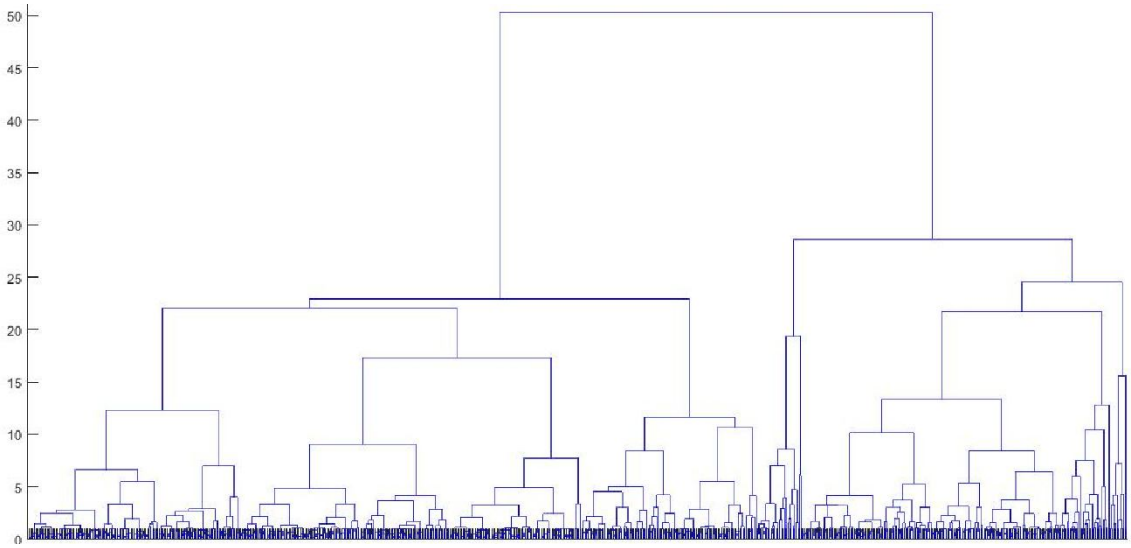

Figura 7: Dendograma da relação dos 7 indicadores, por 660 CNAEs (Figura 4).

Segundo Hair (2009), não há uma regra objetiva para obter a solução final de agrupamentos. Deve-se, porém, evitar agrupamentos pequenos, e buscar obter agrupamentos significativamente diferentes no conjunto de variáveis.

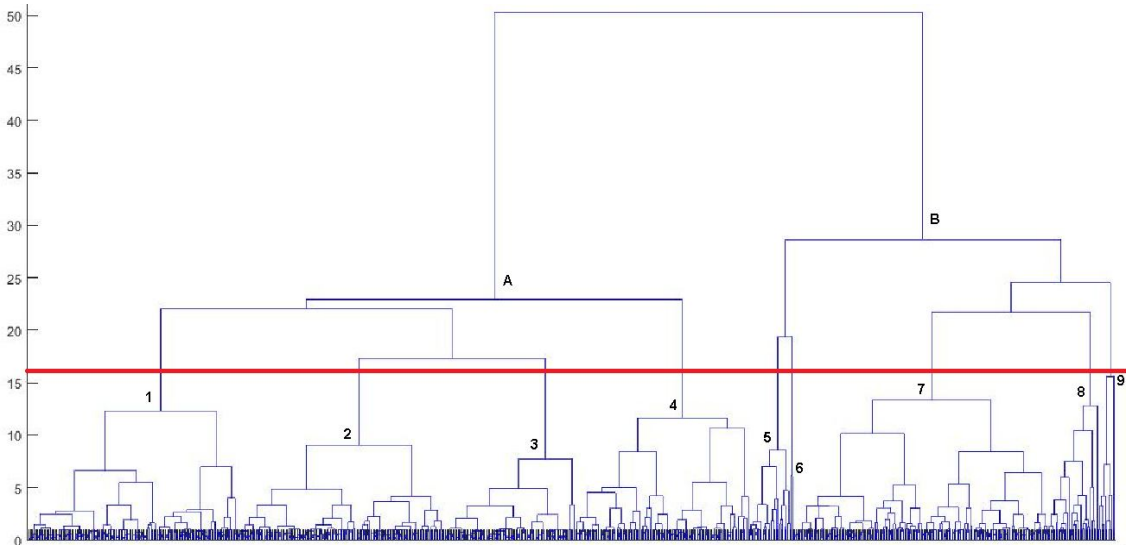

Figura 8: Corte do dendograma e identificação dos agrupamentos.

Seguindo o diagrama de decisão da análise de agrupamento (Figura 2), no estágio de interpretação é necessário examinar os centroides e nomear agrupamentos com base nas suas variáveis. Escolheu-se, portanto, o ponto de corte de, aproximadamente, 16 pontos de similaridade, visto que um corte igual ou abaixo de 15 resultaria em um cluster contendo apenas 1 CNAE contrariando a indicação do parágrafo anterior. Desta forma, obteve-se 9 clusters, representados na Figura 8. Inicialmente, os *clusters* foram numerados de 1 à 9 e identificou-se na hierarquia os *clusters* 1, 2, 3 e 4 pertencentes a um agrupamento maior chamado A, e os demais *clusters* ao agrupamento B, tendo início a análise dos CNAEs que compõem cada *cluster*. Essa identificação permite analisar as características de similaridade entre as atividades de um mesmo *cluster*, e depois realizar suposições em relação às diferenças e semelhanças entre os membros de um *cluster*, e entre *clusters* próximos e distantes. Como critério de corte os CNAEs foram agrupados de acordo com a atividade econômica denominados G_CNAE, indo de 0111 a 9900, conforme Tabela 2:

Tabela 2: Classificação de CNAEs por grupos G_CNAEs.

G_CNAE	DESCRIÇÃO	INTERVALO
A	AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQÜICULTURA	111-322
B	INDÚSTRIAS EXTRATIVAS	500-990
C	INDÚSTRIAS DE TRANSFORMAÇÃO	1011-3329
D	ELETRICIDADE E GÁS	3511-3530
E	ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO	3600-3900
F	CONSTRUÇÃO	4110-4399
G	COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS	4511-4790
H	TRANSPORTE, ARMAZENAGEM E CORREIO	4911-5320
I	ALOJAMENTO E ALIMENTAÇÃO	5510-5620
J	INFORMAÇÃO E COMUNICAÇÃO	5811-6399
K	ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS	6410-6630
L	ATIVIDADES IMOBILIÁRIAS	6810-6822
M	ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS	6911-7500
N	ATIVIDADES ADMINISTRATIVAS E SERVIÇOS COMPLEMENTARES	7711-8299
O	ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL	8411-8430
P	EDUCAÇÃO	8511-8599
Q	SAÚDE HUMANA E SERVIÇOS SOCIAIS	8610-8800
R	ARTES, CULTURA, ESPORTE E RECREAÇÃO	9001-9329
S	OUTRAS ATIVIDADES DE SERVIÇOS	9411-9609
T	SERVIÇOS DOMÉSTICOS	9700
U	ORGANISMOS INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS	9900

FONTE: CONCLA-IBGE (2017)

A Tabela 3 representa a média dos índices por CNAEs para cada *cluster* de forma a identificar quais taxas são mais elevadas dentro dos *clusters* formados.

Tabela 3: Média das taxas por *Clusters*.

Médias	Incidência	Incidência de Doenças Ocupacionais	Incidência de Acidentes Típicos	Incidência de Incapacidade Temporária	Taxa de Mortalidade	Taxa de Letalidade	Acidentalidade para a faixa 16 a 34 anos
Cluster 1	8,3906364	0,2365821	4,4948204	7,1242224	0,6719762	0,8973938	42,953765
Cluster 2	18,43132	0,304402	12,083137	15,585133	2,7810581	1,7990124	58,336097
Cluster 3	6,5944756	0,0982762	2,9270544	6,2637202	1,2172653	1,8313151	67,816775
Cluster 4	12,489888	0,2035602	7,0105799	11,378556	13,658828	12,155563	49,662638
Cluster 5	25,204312	0,8609945	17,676407	21,682051	42,559005	17,600419	51,868037
Cluster 6	8,2972599	0,5027231	5,0283058	8,2972599	75,426379	88,461538	36,923077
Cluster 7	29,501797	0,6444516	21,22299	24,624064	10,102374	3,6683848	54,050126
Cluster 8	52,186821	0,5991678	41,419586	40,891258	13,004108	2,4072712	52,904105
Cluster 9	37,175307	5,828905	20,025109	26,674447	0,9863499	0,3641881	52,905465

Legenda:

	Maiores taxas		Menores taxas

As Figuras 9 a 17 apresentam individualmente cada cluster, e sua composição por grandes grupos de atividades econômicas (G_CNAEs) identificando o G_CNAE mais representativo dentro de cada *cluster*.

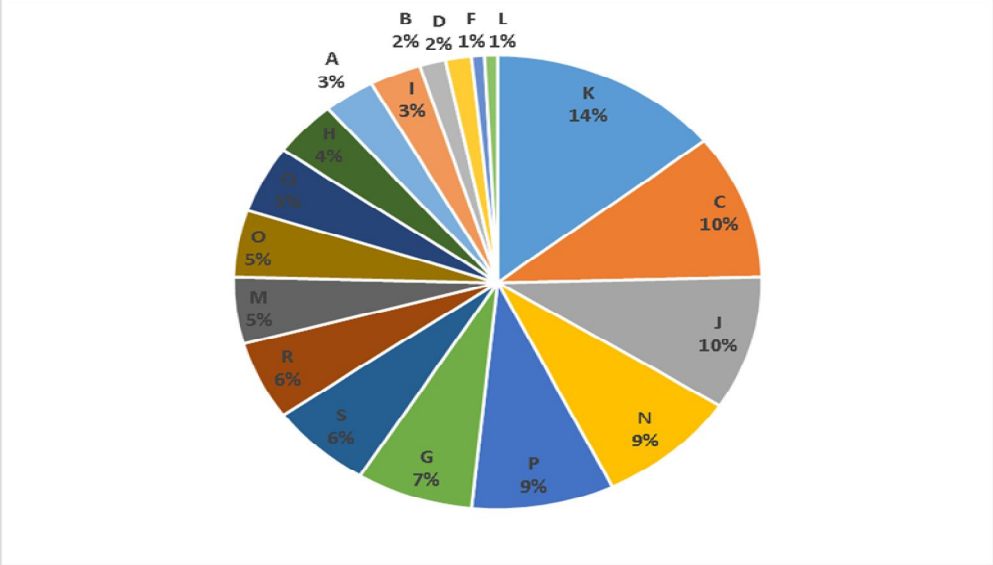

Figura 9: Grupos de CNAE que compõe o CLUSTER 1.

Os *clusters* 1 e 4 (Figuras 9 e 12) são bem heterogêneos, ou seja, compostos por quase todos os G_CNAEs. O esperado desses clusters é que fossem formados predominantemente por um ou poucos grupos (como ocorreu com os demais clusters encontrados), indicando que aquele agrupamento possui características em comum. Entretanto, essa variedade de G_CNAEs indica que há similaridades comuns a quase todos os grupos em relação aos sete

indicadores analisados, como a Taxa de Incidência, Incidência de Doenças Ocupacionais, Incidência de Acidentes Típicos, Incidência de Incapacidade Temporária, Taxa de Mortalidade, Taxa de Letalidade e Acidentalidade para a faixa 16 a 34 anos, requerendo uma investigação mais aprofundada por CNAE para esses grupos.

O *cluster* 1 (Figura 9) possui 127 CNAEs similares e dos 21 G_CNAEs, não estão representados nesse grupo apenas o E, T e U. O G_CNAE com maior representatividade é o K que engloba atividades financeiras, de seguros e serviços relacionados (Tabela 2). Na média das taxas individuais, esse grupo apresenta o menor índice na taxa de mortalidade (0,67) e o segundo menor nas taxas de letalidade (0,89), acidentalidade para a faixa de 16 a 34 anos (42,95), acidentes típicos (4,49) e incapacidade temporária (7,12) (Tabela 3).

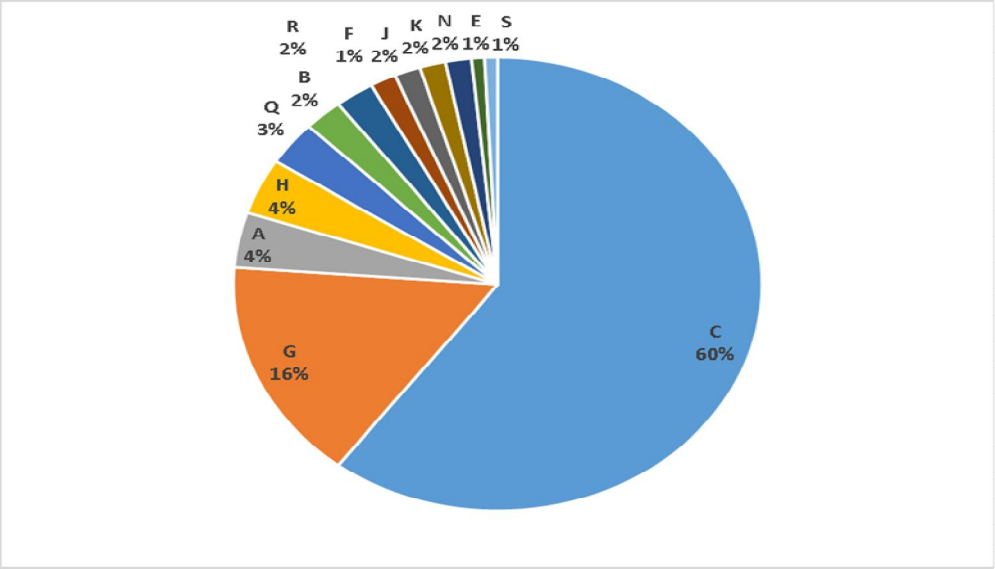

Figura 10: Grupos de CNAE que compõe o CLUSTER 2.

O *cluster* 2 (Figura 10), é composto por 126 CNAEs, predominantemente pelo G_CNAE C, que engloba atividades da indústria de transformação, porém o G_CNAE G, possui a segunda maior representatividade e engloba atividades do comércio; reparação de veículos automotores e motocicletas (Tabela 2). Não estão presentes nesse grupo os G_CNAEs: D; I; L; M; O; P; T e U. Na média das taxas individuais, o grupo não apresenta índices baixos ou muito altos, a única taxa com destaque é acidentalidade para a faixa de 16 a 34 anos (58,33), a segunda maior dentro dos nove *clusters* (Tabela 3).

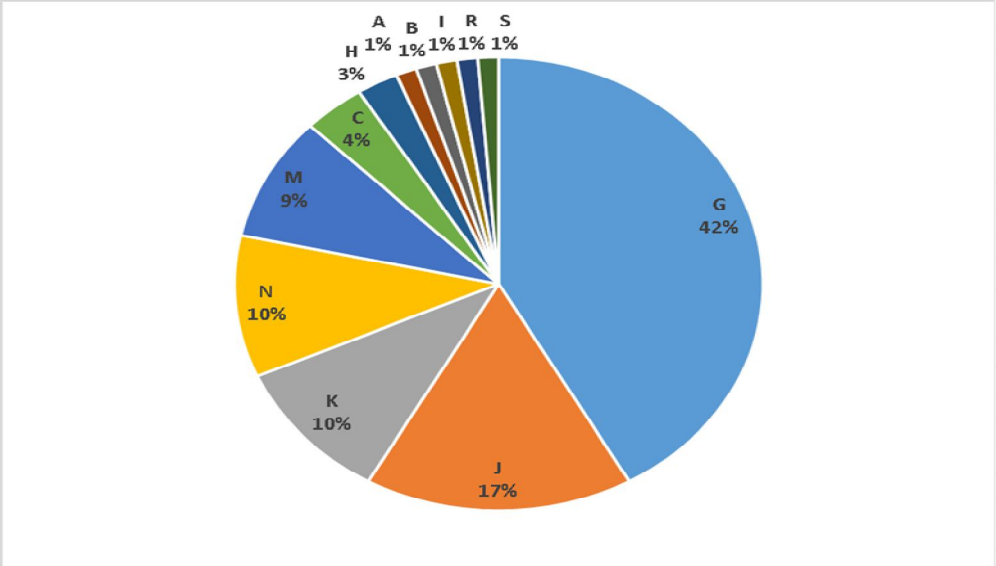

Figura 11: Grupos de CNAE que compõe o CLUSTER 3.

O *cluster* 3 (Figura 11), possui 79 CNAEs, predomina as atividades do G_CNAE G: comércio; reparação de veículos automotores e motocicletas. O segundo grupo com maior representatividade pertence ao G_CNAE J: informação e comunicação (Tabela 2). Não estão representados nesse grupo os G_CNAEs: D; E; F; L; O; P; Q; T e U. Na média das taxas individuais, esse grupo apresenta os menores índices nas taxas de incidência (6,59), incidência de doenças ocupacionais (0,09), incidência de acidentes típicos (2,92), incidência de incapacidade temporária (6,26), e a maior taxa para acidentalidade para a faixa 16 a 34 anos (67,81) (Tabela 3).

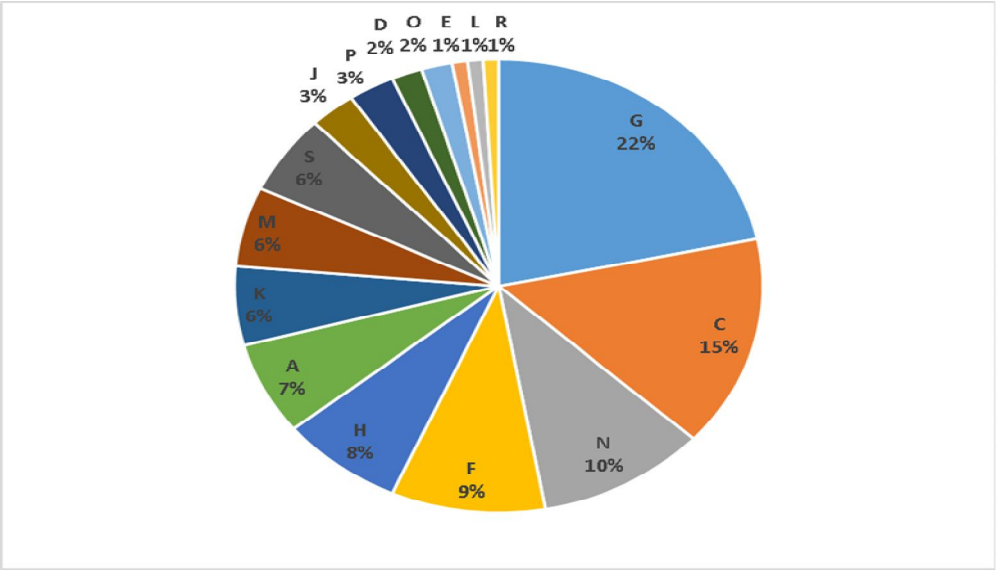

Figura 12: Grupos de CNAE que compõe o CLUSTER 4.

O *cluster* 4 (Figura 12), possui 106 CNAEs, e assim como o *cluster* 1 (Figura 9) é representando pelos G_CNAEs de forma quase heterogenia, o que difere são os grupos de G_CNAEs não representados: B: I: Q; T e U e a predominância da atividade G: comércio; reparação de veículos automotores e motocicletas (Tabela 2). Na média das taxas individuais, esse grupo apresenta o segundo menor índice na taxa incidência de doenças ocupacionais (0,20) (Tabela 3).

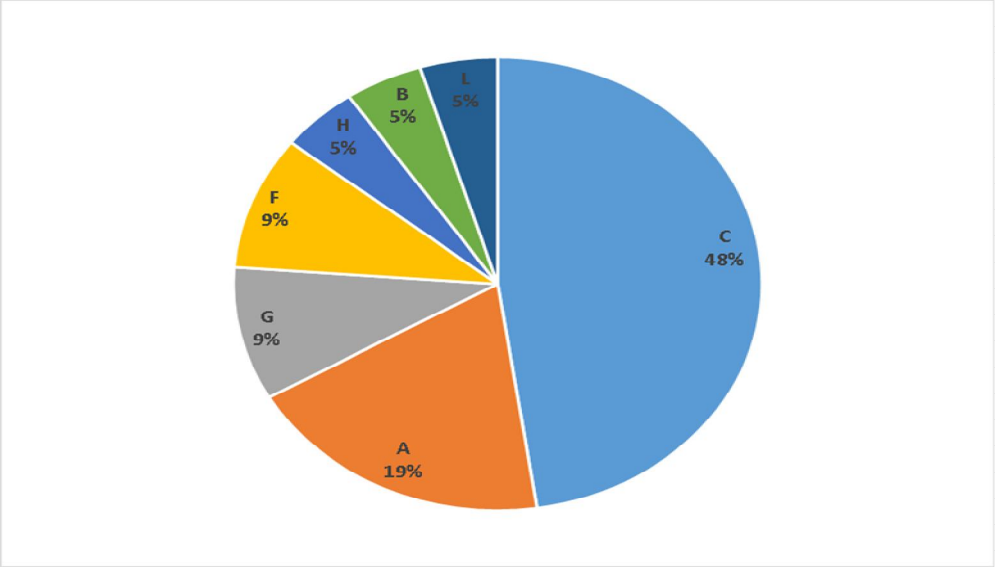

Figura 13: Grupos de CNAE que compõe o CLUSTER 5.

O *cluster* 5 (Figura 13) é um grupo representado por apenas 25 CNAEs e 7 G_CNAEs, tem como maior representante a atividade C (indústrias de transformação) e a segunda maior, a atividade A (agricultura, pecuária, produção florestal, pesca e aquicultura), juntas representam 67% do *cluster*. Na média das taxas individuais, esse grupo apresenta o segundo maior índice nas taxas de mortalidade (42,55) e letalidade (17,60) (Tabela 3).

O *cluster* 6 (Figura 14), só possui 2 CNAEs e 2 G_CNAEs, o A que são as atividades de agricultura, pecuária, produção florestal, pesca e aquicultura e o H das atividades de transporte, armazenagem e correio (Tabela 2). Na média das taxas individuais, esse grupo apresenta o maior índice nas taxas de mortalidade (75,42) e letalidade (88,46) e a menor taxa para acidentalidade para a faixa 16 a 34 anos (36,92) (Tabela 3).

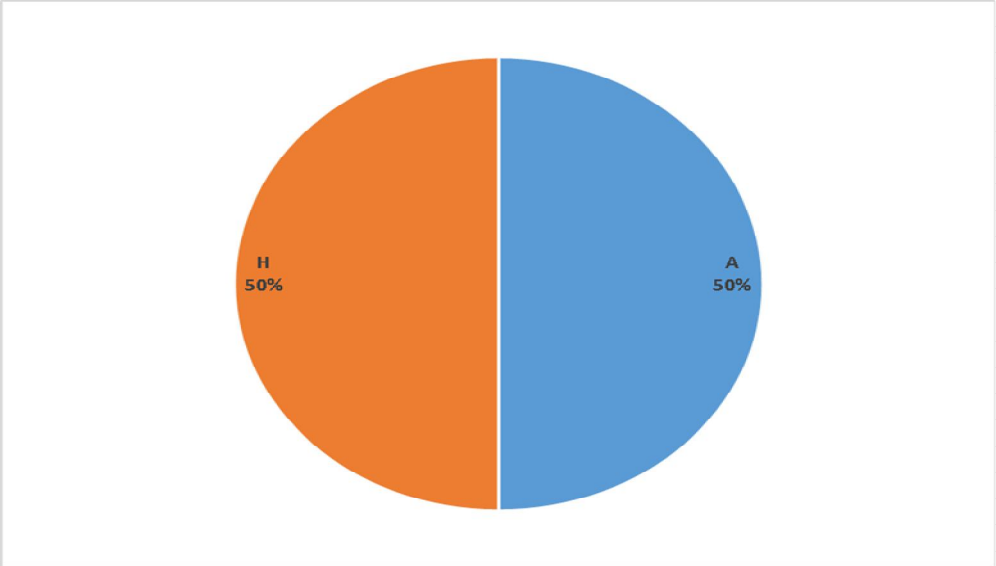

Figura 14: Grupos de CNAE que compõe o CLUSTER 6.

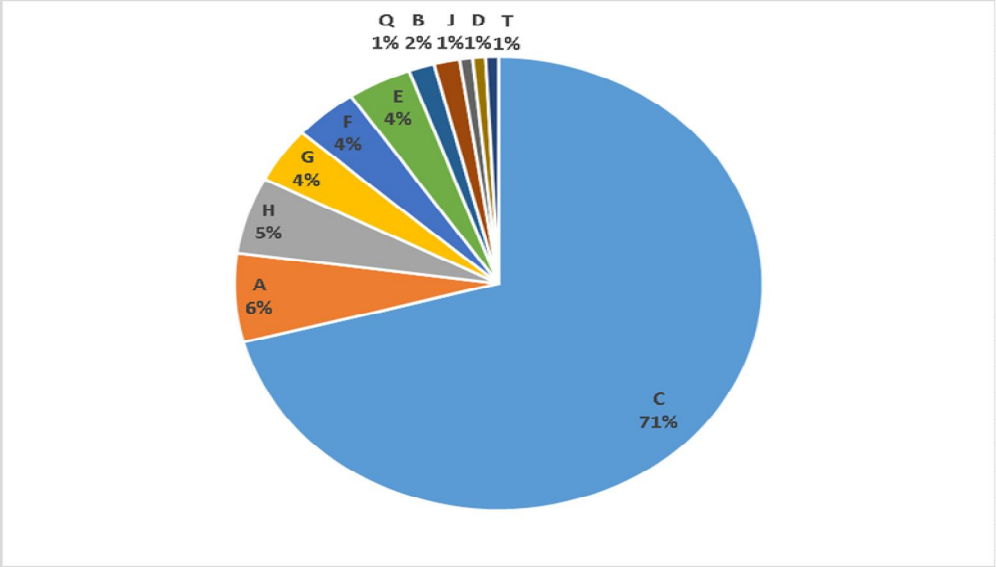

Figura 15: Grupos de CNAE que compõe o CLUSTER 7.

O *cluster* 7 (Figura 15) possui 155 CNAEs, sendo 71% deles pertencentes ao G_CNAE C, indústrias de transformação (Tabela 2). O G_CNAE A que engloba atividades de agricultura, pecuária, produção florestal, pesca e aquicultura é o segundo maior, mas possui 6% de representatividade. O *cluster* 7 é o maior dentre os 9 *clusters* em quantidade de CNAEs similares. Na média das taxas individuais, esse grupo apresenta o segundo maior índice na taxa de incidência de acidentes típicos (21,22) (Tabela 3).

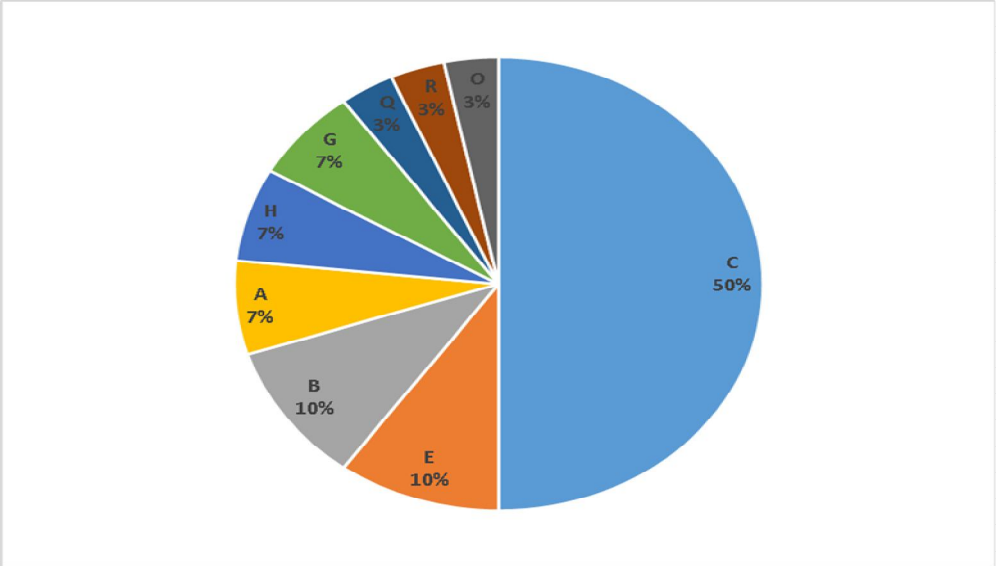

Figura 16: Grupos de CNAE que compõe o CLUSTER 8.

O *cluster* 8 (Figura 16), possui 30 CNAEs que estão representados por 9 G_CNAEs, o maior deles o C das atividades das indústrias de transformação com 50% de representatividade, seguidos do B e E que representam 10% cada um, nesse *cluster* e englobam as atividades de indústrias extrativas (B) e água, esgoto, atividades de gestão de resíduos e descontaminação (E) (Tabela 2). Na média das taxas individuais, esse grupo apresenta o maior índice nas taxas de incidência (52,18), incidência de acidentes típicos (41,41), incidência de incapacidade temporária (40,89) (Tabela 3).

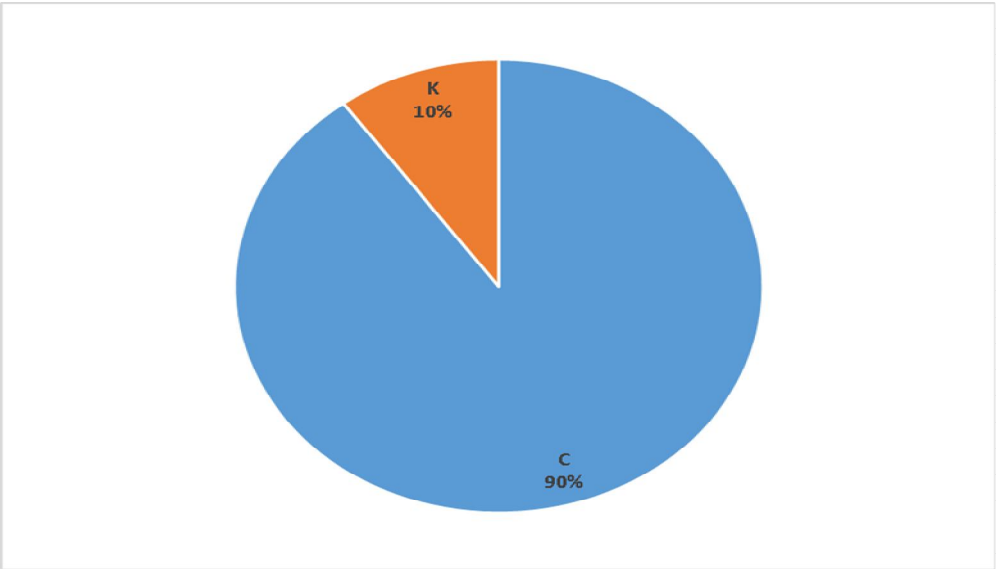

Figura 17: Grupos de CNAE que compõe o CLUSTER 9.

O *cluster* 9 (Figura 17), formado por 10 CNAEs, representados por 2 G_CNAEs C com 90% (indústrias de transformação) e K com 10% (atividades financeiras, de seguros e

serviços relacionados). Na média das taxas individuais, esse grupo apresenta o maior índice na incidência de doenças ocupacionais (5,82) e o segundo maior índice nas taxas de incidência (37,17), incidência de acidentes típicos (20,02), incidência de incapacidade temporária (26,67) e o menor índice na taxa de letalidade (0,36) (Tabela 3).

O grupo A (Figura 8) composto pelos *clusters* 1, 2, 3 e 4 detêm 438 CNAEs e uma similaridade entre os G_CNAEs mais heterogenia, porem os *clusters* 2 e 3 se mostraram similares com atividades predominantes distintas, sendo C para o 2 (engloba atividades da indústria de transformação) e G para o 3 (engloba atividades do comércio; reparação de veículos automotores e motocicletas), esse grupo possui os *clusters* com resultados mais baixos, sendo que o *cluster* 3 possui as menores taxas em quatro dos sete indicadores analisados, mas na taxa que mede a acidentalidade para a faixa 16 a 34 anos, possui o maior resultado (Tabela 3).

Os *clusters* 5, 6, 7, 8 e 9 que pertencem ao grupo B (Figura 8) possuem menos CNAEs, exceção ao *cluster* 7 que é o maior em CNAEs (155) e representam em seus agrupamentos a predominância da atividade G_CNAE C, e o *cluster* 6, formado por apenas dois elementos e com taxas muito elevadas nos indicadores de mortalidade e letalidade (Tabela 3). Apesar desse grupo apresentar taxas maiores que o grupo A, o *cluster* 6 têm a menor taxa para acidentalidade para a faixa 16 a 34 anos e o *cluster* 9 têm a menor taxa para a letalidade (Tabela 3). Os *clusters* que compõem o grupo A (1, 2, 3 e 4) apresentam, em sua maioria, os menores índices, na Tabela 4 estão dispostos os cinco melhores resultados por incidência com a descrição da atividade econômica.

As Tabelas 4 e 5 trazem a descrição das atividades econômicas com melhores e piores índices. Entender o funcionamento destas atividades é fundamental para descobrir as similaridades na execução dessas atividades e das ocupações que as compõem, orientando quais medidas devem ser adotadas. A partir dos dados apresentados nas tabelas 4 e 5, constatou-se que a distribuição de G_CNAEs, tanto para as atividades econômicas com menores incidências de acidente de trabalho quanto para as atividades econômicas com taxas de incidência mais elevadas, não se restringiu a um G_CNAE específico, englobando desde atividades de cultivo, criação de animais, transporte, fabricação, fundição, entre outros (para o grupo B) e atividades de comércio, educação, financeiras, publicidade entre outras para o grupo A.

Tabela 4: As 5 atividades econômicas com menores índices de acidentalidade por *clusters* no grupo A.

CNAE	Descrição	1	2	3	4	5	6	7	G_CNAE
Cluster 1									
9492	Atividades de organizações políticas	1,12	0,00	0,56	0,94	0,00	0,00	33,33	S
6450	Sociedades de capitalização	1,98	0,00	0,00	1,98	0,00	0,00	50,00	K
7311	Agências de publicidade	2,07	0,06	0,57	2,10	0,00	0,00	56,52	M
6010	Atividades de rádio	2,13	0,13	0,79	2,03	0,00	0,00	38,27	J
8593	Ensino de idiomas	2,16	0,03	0,68	2,13	1,44	6,67	56,00	P
Cluster 2									
4724	Comércio varejista de hortifrutigranjeiros	9,49	0,02	5,43	9,35	3,54	3,72	59,59	G
4511	Comércio a varejo e por atacado de veículos automotores	9,52	0,11	5,20	9,10	4,52	4,75	59,23	G
9329	Atividades de recreação e lazer não especificadas anteriormente	9,97	0,24	5,89	9,09	6,81	6,83	54,95	R
9101	Atividades de bibliotecas e arquivos	10,53	0,44	8,77	10,53	0,00	0,00	79,17	R
1821	Serviços de pré-impressão	10,84	0,00	6,03	10,23	6,09	5,62	57,87	C
Cluster 3									
6492	Securitização de créditos	1,17	0,00	0,00	1,17	0,00	0,00	100,00	K
5912	Atividades de pós-produção cinematográfica, de vídeos e de programas de televisão	1,92	0,64	0,00	1,92	0,00	0,00	66,67	J
6319	Portais, provedores de conteúdo e outros serviços de informação na internet	2,25	0,00	0,52	2,13	0,00	0,00	74,36	J
6440	Arrendamento mercantil	2,45	0,00	2,45	0,00	0,00	0,00	100,00	K
6391	Agências de notícias	2,81	0,00	1,69	2,81	0,00	0,00	60,00	J
Cluster 4									
6424	Crédito cooperativo	2,91	0,12	0,60	2,72	4,12	14,18	63,12	K
4755	Comércio varejista especializado de tecidos e artigos de cama, mesa e banho	3,98	0,09	1,60	3,90	4,75	11,93	58,45	G
8542	Educação profissional de nível tecnológico	4,12	0,40	2,79	3,99	13,29	32,26	41,94	P
7420	Atividades fotográficas e similares	4,39	0,11	2,17	3,91	5,29	12,05	61,45	M
4756	Comércio varejista especializado de instrumentos musicais e acessórios	4,51	0,00	1,41	4,37	14,09	31,25	56,25	G
Legenda									
1 = Incidência 2 = Incidência de Doenças Ocupacionais 3 = Incidência de Acidentes Típicos 4 = Incidência de Incapacidade Temporária 5 = Taxa de Mortalidade 6 = Taxa de Letalidade 7 = Acidentalidade para a faixa 16 a 34 anos									

O que distingue os CNAEs do *cluster* 1 para o *cluster* 3 são as taxas para acidentalidade para a faixa 16 a 34 anos, que no *cluster* 3 são mais elevadas.

Os *clusters* que compõem o grupo B (5, 6, 7, 8 e 9) apresentam, em sua maioria, os maiores índices, na Tabela 5 estão dispostos os cinco piores resultados por incidência com a descrição da atividade econômica, com exceção para o *cluster* 6 que possui apenas 2 CNAEs.

Tabela 5: As 5 atividades econômicas com maiores índices para acidentalidade por *clusters* do grupo B.

CNAE	Descrição	1	2	3	4	5	6	7	G_CNAE
Cluster 5									
132	Cultivo de uva	30,96	0,07	19,97	30,76	71,54	23,11	47,48	A
2920	Fabricação de caminhões e ônibus	31,59	3,56	21,22	19,42	18,71	5,92	44,43	C
1932	Fabricação de biocombustíveis, exceto álcool	33,50	0,00	26,36	30,09	62,03	18,52	69,44	C
3315	Manutenção e reparação de veículos ferroviários	33,50	3,14	22,51	28,27	52,35	15,63	57,81	C
2443	Metalurgia do cobre	36,79	2,98	26,58	27,64	21,26	5,78	46,82	C
Cluster 6									
5022	Transporte por navegação interior de passageiros em linhas regulares	6,54	0,00	3,02	6,54	50,31	76,92	53,85	H
153	Criação de caprinos e ovinos	10,05	1,01	7,04	10,05	100,54	100,00	20,00	A
Cluster 7									
1910	Coquerias	40,78	0,00	21,36	34,95	0,00	0,00	66,67	C
4940	Transporte dutoviário	41,12	0,31	31,42	15,17	0,00	0,00	34,98	H
2531	Produção de forjados de aço e de metais não-ferrosos e suas ligas	42,32	2,53	32,19	38,94	0,00	0,00	53,49	C
2424	Produção de relaminados, trefilados e perfilados de aço	44,55	2,55	34,42	36,19	7,08	1,59	55,64	C
2431	Produção de tubos de aço com costura	46,52	1,73	36,51	36,87	14,40	3,10	58,20	C
Cluster 8									
2539	Serviços de usinagem, solda, tratamento e revestimento em metais	63,69	1,48	45,63	58,96	9,85	1,55	61,27	C
725	Extração de minerais radioativos	68,11	0,00	55,73	22,70	0,00	0,00	48,48	B
141	Produção de sementes certificadas	68,86	0,27	56,92	66,65	73,83	10,72	39,57	A
2451	Fundição de ferro e aço	82,01	2,24	69,05	45,50	3,99	0,49	58,33	C
5310	Atividades de Correio	109,20	4,71	83,08	92,83	10,33	0,95	31,60	H
Cluster 9									
2851	Fabricação de máquinas e equipamentos para a prospecção e extração de petróleo	43,42	17,01	20,77	25,57	0,00	0,00	69,23	C
2942	Fabricação de peças e acessórios para os sistemas de marcha e transmissão de veículos automotores	46,63	4,36	32,17	30,79	0,00	0,00	57,17	C
3031	Fabricação de locomotivas, vagões e outros materiais rodantes	52,29	7,44	29,21	39,63	0,00	0,00	53,02	C
3091	Fabricação de motocicletas	52,91	3,68	32,61	36,02	0,00	0,00	59,76	C
2550	Fabricação de equipamento bélico pesado, armas de fogo e munições	53,60	3,67	36,80	42,45	0,00	0,00	43,87	C
Legenda									
1 = Incidência 2 = Incidência de Doenças Ocupacionais 3 = Incidência de Acidentes Típicos 4 = Incidência de Incapacidade Temporária 5 = Taxa de Mortalidade 6 = Taxa de Letalidade 7 = Acidentalidade para a faixa 16 a 34 anos									

Num ranking geral por incidência, destacaram-se negativamente as atividades de correios (*cluster* 8) como a de maior taxa de acidentalidade, seguida da fundição de ferro e aço (*cluster* 8). Os melhores resultados gerais (menor acidentalidade) foram atribuídos às atividades de organizações políticas (*cluster* 1), seguida pela atividade de securitização de créditos (*cluster* 3).

5 CONCLUSÃO

As técnicas de data mining permitem analisar grande quantidade de dados, encontrando padrões e formando agrupamentos. É vantajosa pois o problema de pesquisa não precisar ser muito específico: escolhe-se a questão de investigação, coleta-se e formata-se os dados a serem minerados, e então analisa-se posteriormente os agrupamentos formados pela aplicação dos diversos métodos, até encontrar o modelo mais próximo da realidade segundo as indicações sugeridas pelo *clustering*.

Embora haja ferramentas computacionais que realizem a análise desses dados e a geração automática de dendograma, a ausência de critérios de corte específicos, bem como a necessidade de conhecimentos em programação para aplicação da técnica, torna-a pouco acessível. Entretanto, essa dificuldade pode ser minimizada por meio de pesquisas com equipe multidisciplinar, onde um profissional de *data mining* realiza a mineração e um profissional da área investigada propõe os critérios de corte, a escolha da melhor representação gráfica, e por fim, a interpretação dos resultados.

Os acidentes aqui investigados são apresentados por CNAE, pertencentes a um G_CNAE, e ações preventivas tendem a utilizar essas classificações para selecionar o público alvo dessas ações. Entretanto, os clusters aqui encontrados indicam que os 7 indicadores avaliados são similares a diversos grupos em alguns clusters, sugerindo ações globais por parte do governo, e não setorializadas para um ramo específico. No *cluster* 3 observou-se baixo índice nas taxas de incidência dos acidentes, onde pode-se inferir que esse grupo possui o menor risco à acidentes. Já o *cluster* 6, com altas taxas de mortalidade e letalidade, e composto por apenas 2 CNAEs, infere-se um risco alto de acidentes. Para este *cluster*, recomenda-se uma investigação mais minuciosa, a fim de identificar as práticas ou falhas dos programas de SST que podem estar relacionadas a essas taxas elevadas. O *cluster* 8 também indica alto risco de acidentes, pois apresenta as maiores taxas para: incidência, incidência de acidentes típicos e incidência de incapacidade temporária.

Apesar dos CNAEs serem diferentes, entre si, quando orientados ao tipo de atividade, eles também apresentam similaridade quanto à multifuncionalidade das empresas, onde uma função está presente em diversos segmentos. Essa dificuldade em analisar os índices por G_CNAE foi percebida observando que os tipos de empresa que se classificam como indústria da transformação (C) estão atrelados a diversos tipos de manipulação de materiais, desde beneficiamento de madeira, metal e químicas, podendo esse, ser o motivo pelo qual o G_CNAE C está presente em quase todos os *clusters*. Também o fato do CNAE não

representar uma única atividade laboral nas empresas, podendo ser resolvido aplicando esta técnica aos acidentes de trabalho registrados por ocupações.

Notou-se que os *clusters* que formam o grupo A possuem tamanhos próximos e representam a maior parte dos CNAEs e G_CNAEs, o que sugere estudos futuros que realizem um aprofundamento nesse grupo. Concomitante a isso, é possível ainda realizar um estudo temporal para verificar se o *clusters* formados com esses dados se manterão com os mesmos CNAEs para anos anteriores, verificando, além da similaridade, possíveis tendências.

O uso do software Matlab permitiu a aplicação do método Ward e a geração do dendograma, e o uso do Microsoft Excel permitiu a tabulação dos dados e a representação gráfica dos clusters. Entretanto, outros softwares tais como o Iramuteq, software que executa sobre o R, e o RStudio, podem ser utilizados para a geração de dendogramas, e criação de bancos de dados.

Por fim, a aplicação da técnica foi alcançada, embora a análise dos *clusters* formados e suas implicações não foram esgotadas e podem ser avaliadas em estudos futuros, mantendo ou alterando o método selecionado, bem como a linha de corte utilizada para formar os agrupamentos, pois diferentes medidas de similaridade e outros algoritmos podem afetar o resultado obtido. A análise de agrupamentos mostrou-se muito útil na identificação de padrões não percebidos por outras técnicas multivariadas.

REFERÊNCIAS

Anuário Estatístico de Acidentes do Trabalho : AEAT 2014 / Ministério do Trabalho e Previdência Social ... [et al.]. – vol. 1 – . – Brasília : MTPS, 2014. 990 p. Disponível em: http://www.previdencia.gov.br/dados-abertos/dados-abertos-previdencia-social/

CASTRO, Manuel Altino Torres Aniceto. **Projecto Agrupamento – "Clustering"**. ISEP - Instituto Superior de Engenharia do Porto, 2003.

CONSTANTINO, André F. et al. **Análise do teor e da qualidade dos lipídeos presentes em sementes de oleaginosas por RMN de baixo campo**. Química Nova, v. 37, n.1, 10-17, 2014.

CÔRTES, Sérgio da Costa. PORCARO, Rosa Maria. LIFSCHITZ, Sérgio. **Mineração de dados – funcionalidades, técnicas e abordagens**. PUC-RioInf.MCC10/02, 2002.

HAIR, J. F., et al. **Análise multivariada de dados**. Trad. Adonai S. Sant'Anna. 6 ed. Porto Alegre: Bookman, 2009.

HEUSER, Carlos Alberto. **Projeto de Banco de Dados**. 6ª ed. Bookman, Porto Alegre, p.20-25, 2009.

Instituto Brasileiro de Geografia e Estatística – IBGE. Disponível em: http://concla.ibge.gov.br/

KUBRUSLY, L. S. Um procedimento para calcular índices a partir de uma base de dados multivariados. Pesquisa Operacional, Rio de Janeiro, v.21, n. 1, 2001.

LIU, Bing. **Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data**. 2ª ed. Springer, New York, 2011.

MALHOTRA, N. Pesquisa de marketing: uma orientação aplicada. Trad. Laura Bocco. 4 ed. Porto Alegre: Bookman, 2006.

MOREIRA, Evaristo. **Como criar um sistema de recomendação**, 2015. Disponível em: http://evaristomoreira.com.br/artigos/como-criar-um-sistema-de-recomendacao/#comment-271

PASTORE, José. **O custo dos acidentes e doenças do trabalho no Brasil**, 2012. Disponível em: http://economia.estadao.com.br/noticias/geral,pais-gasta-r-72-bilhoes-por-ano-com-acidente-de-trabalho-imp-,825342

REZENDE, Solange Oliveira, Org. **Sistemas inteligentes: fundamentos e aplicações**. Manole, Barueri – SP, p.8, 2005.

SEIDEL, Enio Junior et al. **Comparação entre o método de ward e o método k-médias no agrupamento de produtores de leite**. Ciência e Natura, UFSM, 30 (1): 7-15, 2008.

SPERANDIO, Carlos A. **Fundamentos de engenharia de segurança:** cursos de engenharia: mecânica, eletrônica e eletrotécnica. Curitiba: CEFET-PR. 1998. 253 p

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. 2006. **Introduction to Data Mining**. Pearson Education, Inc. Boston USA.

WARD, J. H. **Hierarquical grouping to optimize an objective function.** Journal of the American Statistical Association, v. 58, p. 236 – 244. Mar. 1963.