

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Danilo Arantes da Silva

**Aplicação de técnicas de pré-processamento e
agrupamento na base de dados de benefícios
previdenciários do Ministério Público do
Trabalho**

Uberlândia, Brasil

2018

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Danilo Arantes da Silva

**Aplicação de técnicas de pré-processamento e
agrupamento na base de dados de benefícios
previdenciários do Ministério Público do Trabalho**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, Minas Gerais, como
requisito exigido parcial à obtenção do grau
de Bacharel em Sistemas de Informação.

Orientador: Elaine Ribeiro de Faria Paiva

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2018

Danilo Arantes da Silva

Aplicação de técnicas de pré-processamento e agrupamento na base de dados de benefícios previdenciários do Ministério Público do Trabalho

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Uberlândia, Brasil, 2 de julho de 2018:

Elaine Ribeiro de Faria Paiva
Orientador

Maurício Cunha Escarpinati

Paulo Henrique Ribeiro Gabriel

Uberlândia, Brasil
2018

Agradecimentos

Agradeço primeiramente a Deus por me dar saúde, força e ânimo para concluir mais esta etapa da minha vida.

Aos meus pais pelo carinho, amor e educação que me deram e por sempre me incentivarem a estudar.

Ao meu irmão, pois mesmo reclamando, contribuiu nas horas que eu precisava de concentração e silêncio para estudar.

À minha namorada, pois sem ela nada disso estaria acontecendo. Obrigado por todo o apoio, pela paciência, por entender quando não podíamos nos ver devido a trabalhos e provas e por ser essa pessoa especial em minha vida.

À minha orientadora pelo acompanhamento no decorrer do projeto, pela paciência e pelo dom de ensinar.

E por fim aos meus familiares e amigos por tornarem meus dias mais alegres e por sempre estarem comigo.

Resumo

A grande quantidade de acidentes e doenças de trabalhadores e a quantidade de dinheiro que são desembolsados com benefícios acidentários no Brasil são preocupantes. Diante disso, o Ministério Público do Trabalho juntamente com a Organização Internacional do Trabalho lançaram em abril de 2017 o Observatório Digital de Saúde e Segurança do Trabalho, a fim de facilitar o acesso a estatísticas sobre acidentes de trabalho e benefícios concedidos a trabalhadores que, antes se encontravam perdidas em banco de dados governamentais. Nesse observatório foi disponibilizado duas bases de dados: uma de acidentes de trabalhos notificados e outra de benefícios previdenciário concedidos à trabalhadores no Brasil. Este trabalho tem por objetivo analisar a base de dados de benefícios afim de nortear tomadas de decisões de políticas de controle e prevenção de acidentes e doenças ocupacionais. Para isso, foi criado uma ferramenta em Java afim de realizar técnicas de pré-processamento nos dados e também foram utilizadas técnicas de agrupamento nos dados afim de se buscar padrões nos mesmos. A partir da base de dados de benefício disponibilizada foram criadas várias outras bases a fim de explorar as técnicas de agrupamento em diferentes visões do problema. Os resultados do agrupamento foram avaliados usando a medida de silhueta simplificada e indicam que os algoritmos utilizados não mostraram bom desempenho na base de dados, em especial, devido a alta dimensionalidade da base após o pré-processamento. Novos algoritmos devem ser explorados, assim como novos métodos de pré-processamento.

Palavras-chave: Agrupamento, benefícios previdenciários, acidentes de trabalho, pré-processamento.

Lista de ilustrações

Figura 1 – Tela inicial do Observatório Digital de Saúde e Segurança do Trabalho	18
Figura 2 – Gráficos do Observatório Digital de Saúde e Segurança do Trabalho . .	19
Figura 3 – Etapas do processo de KDD	20
Figura 4 – Dendograma representando agrupamento hierárquico	27
Figura 5 – Exemplo de execução do <i>k-means</i> em uma base com 3 grupos	29
Figura 6 – Exemplo do agrupamento <i>canopy</i>	30
Figura 7 – Conjunto com grupos circulares (conjunto 1), não circulares (conjuntos 2 e 3) e com ruídos (conjunto 3).	32
Figura 8 – Agrupamento DBSCAN para 3.0000 pontos bidimensionais	33
Figura 9 – Exemplo de hierarquia do CNAE.	41
Figura 10 – Exemplo da planilha de Mesorregiões do DATASUS.	42
Figura 11 – Arquivo no formato ARFF.	43
Figura 12 – Tela da ferramenta desenvolvida neste trabalho para pré-processamento dos dados.	44
Figura 13 – Gráficos das bases de dados de doenças utilizando o algoritmo <i>k-means</i> e modificando o número de agrupamento.	49
Figura 14 – Gráficos das bases de dados de acidentes utilizando o algoritmo <i>k-means</i> e modificando o número de agrupamento.	51
Figura 15 – Gráficos das bases de dados de doenças utilizando o algoritmo <i>canopy</i> e modificando o número de agrupamento.	53
Figura 16 – Gráficos das bases de dados de acidentes utilizando o algoritmo <i>canopy</i> e modificando o número de agrupamento.	54
Figura 17 – Gráficos das bases de dados de doenças utilizando o algoritmo EM e modificando o número de agrupamento.	56
Figura 18 – Gráficos das bases de dados de acidentes utilizando o algoritmo EM e modificando o número de agrupamento.	57

Lista de tabelas

Tabela 1 – Descrição dos dados do conjunto de benefícios previdenciários	38
Tabela 2 – Codificação inteira-binária de um atributo categorizado	40
Tabela 3 – Codificação 1-de-n de um atributo categorizado	41
Tabela 4 – Agrupamento <i>k-means</i> , Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.	48
Tabela 5 – Agrupamento <i>k-means</i> , Base_ano_1n_D, onde o ano vai de 2012 à 2016.	49
Tabela 6 – Agrupamento <i>k-means</i> - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.	50
Tabela 7 – Agrupamento <i>k-means</i> - Base_ano_1n_A, onde o ano vai de 2012 à 2016.	50
Tabela 8 – Agrupamento <i>canopy</i> - Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.	52
Tabela 9 – Agrupamento <i>canopy</i> - Base_ano_1n_D, onde o ano vai de 2012 à 2016.	52
Tabela 10 – Agrupamento <i>canopy</i> - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.	53
Tabela 11 – Agrupamento <i>canopy</i> - Base_ano_1n_A, onde o ano vai de 2012 à 2016.	54
Tabela 12 – Agrupamento EM - Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.	55
Tabela 13 – Agrupamento EM - Base_ano_1n_D, onde o ano vai de 2012 à 2016.	55
Tabela 14 – Agrupamento EM - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.	56
Tabela 15 – Agrupamento EM - Base_ano_1n_A, onde o ano vai de 2012 à 2016.	57

Lista de abreviaturas e siglas

MPT	Ministério Público do Trabalho
CAT	Comunicação de Acidente de Trabalho
MD	Mineração de dados
KDD	Knowledge Discovery in Databases
OIT	Organização Internacional do Trabalho
MPU	Ministério Público da União
IA	Inteligência Artificial
EPI	Equipamento de Proteção Individual
INSS	Instituto Nacional do Seguro Social
CNAE	Classificação Nacional de Atividades Econômicas

Sumário

1	INTRODUÇÃO	10
1.1	Justificativa	11
1.2	Objetivos	13
1.2.1	Objetivo geral	13
1.2.2	Objetivos específicos	13
2	REVISÃO BIBLIOGRÁFICA	14
2.1	Introdução	14
2.2	Acidente de Trabalho	14
2.2.1	Comunicação de Acidente de Trabalho - CAT	15
2.3	Benefícios Previdenciários	16
2.3.1	Auxílio-doença acidentário	16
2.3.2	Aposentadoria por invalidez acidentária	16
2.3.3	Pensão por morte por acidente de trabalho	16
2.3.4	Auxílio-acidente	17
2.4	Políticas do Ministério Público do Trabalho para gerenciar acidentes	17
2.5	Descoberta de conhecimento em base de dados	19
2.5.1	Seleção dos dados	19
2.5.2	Pré-Processamento dos dados	20
2.5.3	Transformação dos dados	20
2.5.4	Mineração de Dados	21
2.5.5	Avaliação dos resultados	21
2.6	Pré-processamento	21
2.6.1	Qualidade dos dados	21
2.6.2	Agregação	23
2.6.3	Amostragem	23
2.6.4	Redução de dimensionalidade	23
2.6.5	Seleção de subconjuntos de atributos	23
2.6.6	Criação de atributos	24
2.6.7	Discretização e Binarização	24
2.7	Mineração de Dados	25
2.7.1	Classificação	25
2.7.2	Regressão	25
2.7.3	Regras de Associação	25
2.7.4	Agrupamento	26

2.7.4.1	Métodos Hierárquicos	26
2.7.4.2	Métodos Particionais	27
2.7.4.2.1	<i>K-means</i>	28
2.7.4.2.2	<i>Canopy</i>	28
2.7.4.2.3	<i>Expectation Maximization</i> (EM)	30
2.7.4.3	Métodos baseados em densidade	31
2.7.4.3.1	DBSCAN	32
2.8	Validação de Agrupamento	34
2.9	Trabalhos Relacionados	35
2.10	Considerações Finais	36
3	DESENVOLVIMENTO	37
3.1	Introdução	37
3.2	Apresentação da base de dados do MPT	37
3.3	Pré-processamentos realizados	40
3.4	Bases de dados geradas	44
3.5	Métodos de agrupamento e medida de validação utilizados	46
3.6	Considerações Finais	47
4	RESULTADOS	48
4.1	Introdução	48
4.2	Algoritmo <i>k-means</i>	48
4.3	Algoritmo <i>Canopy</i>	51
4.4	Algoritmo EM	55
4.5	Considerações Finais	57
5	CONCLUSÃO	59
5.1	Contribuições	59
5.2	Trabalhos Futuros	59
	REFERÊNCIAS	61

1 Introdução

Devido ao tamanho da população brasileira e do grande número de empresas existentes, o Brasil está entre os países com maior número de acidentes, doenças e mortes de trabalhadores ([MENDONÇA, 2017](#)). Nos anos de 2012 à 2016, foram registrados cerca de 3,5 milhões de ocorrências de acidentes de trabalho, em torno de R\$ 20 bilhões foram desembolsados com benefícios acidentários e em média 250 milhões de dias de trabalhos foram perdidos ([BRASIL, 2017](#)).

As doenças e acidentes de trabalho causam prejuízos tanto para os trabalhadores e seus familiares quanto para os empregadores. A diferença é que os prejuízos dos trabalhadores são físicos e emocionais enquanto a dos empregadores são financeiros. O sofrimento, a dor e os gastos com a saúde, são exemplos de consequências para os trabalhadores, já os consertos de máquinas danificadas, pagamento de salários a trabalhos não realizados e indenizações, são exemplos de consequências para os empregadores ([GEP/MTSS, 2009](#)).

O Ministério Público do Trabalho (MPT) é um órgão que supervisiona o cumprimento das leis trabalhistas, o relacionamento entre uma organização e sindicatos que os representam, e que procede evitando transtornos nessa área quando houver interesse público. O MPT apresenta um papel fundamental para a sociedade, pois está envolvido em todo tipo de trabalho que esteja vinculado ao trabalhador ([MERELES, 2017](#)).

Todos os acidentes de trabalho ou de trajeto e doença ocupacional devem ser informados à Previdência Social por meio de um documento chamado Comunicação de Acidente de Trabalho (CAT). A CAT deve ser emitida até o primeiro dia útil após o acidente e em caso de óbito deve ser emitida imediatamente. O não cumprimento do prazo de entrega da CAT ou omissão do acontecido ocasiona em multa para a empresa ([INSS, 2017](#)).

Caso o trabalhador acidentado fique incapacitado de realizar suas atividades no trabalho por mais de 15 dias e comprovar por meio de perícia médica, ele pode solicitar o auxílio doença acidentário, que é um benefício previdenciário pecuniário, ou seja, o trabalhador receberá um valor mensalmente pelo INSS por prazo indeterminado ([DUARTE, 2015](#)).

O MPT possui acesso a uma base de dados de CATs registradas e uma base de dados de benefícios previdenciários entre o ano de 2012 a 2016, os quais contém dados importantes para análises sobre os acidentes de trabalho e benefícios concedido a trabalhadores no Brasil. Alguns dos dados da base de benefícios são: município e UF do benefício, idade e sexo do beneficiário, despesa total, descrição da classe do CNAE da empresa empregadora, dentre outros. A análise dos dados da base de benefícios é importante para se compreender os valores gastos com acidentes de trabalho bem como o perfil dos

acidentados. Pode também contribuir para o direcionamento de políticas públicas para o combate e prevenção dos acidentes. São exemplos de análises a serem obtidas a partir desses dados: identificar a faixa etária, analisar qual época do ano e qual região gasta mais com benefícios e quais as atividades econômicas apresentam maiores riscos.

No entanto, realizar a análise manual desses dados é inviável, visto que há uma grande base de benefícios previdenciários já cadastradas, e diariamente novos benefícios são registrados, aumentando assim o volume da base de dados. Em consequência disso, os padrões e até mesmo as anomalias contidas nos dados sofrem evoluções, de modo que buscar essas evoluções exigiria um esforço ainda maior, ocasionando falhas ou impossibilitando o entendimento da movimentação do processo. Em virtude dos fatos mencionados, o desenvolvimento de técnicas de mineração de dados (MD) se faz necessário.

Os métodos de MD são capazes de descobrir automaticamente informações e padrões importantes em grandes volumes de dados que por outros tipos de análises poderiam não serem detectados. Ele é uma das etapas mais importantes do processo de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* ou KDD), que é o processo global de extração de informações a partir de dados (MADEIRA, 2013).

Uma técnica muito importante em MD é a de agrupamento de dados (*clustering*), que tem como objetivo separar objetos em grupos de acordo com suas características. Basicamente a intenção do agrupamento é colocar em um mesmo grupo objetos semelhantes segundo algum critério definido, de forma que as características dos objetos de um mesmo grupo sejam similares entre si e distintas dos objetos de outros grupos (LINDEN, 2009).

Assim, a aplicação de técnicas de agrupamento nos dados de benefícios podem estimar o quanto será gasto nos anos seguintes com benefícios previdenciários, enquanto aplicados nos dados das CATs, podem ajudar a mapear o perfil dos dados trabalhistas no país, facilitando o entendimento das ocorrências dos acidentes. Além disso, os resultados desses agrupamentos podem auxiliar o MPT a reconhecer profissões e empresas que expõe o trabalhador à maiores riscos e direcionar políticas públicas para combater e prevenir estes acidentes.

1.1 Justificativa

Devido à importância dos estudos sobre as condições de trabalho e com a intenção de levar informações perdidas em banco de dados do governo para políticas públicas de prevenção de acidentes e doenças no trabalho, o MPT juntamente com a Organização Internacional do Trabalho (OIT) lançaram em abril de 2017 o Observatório Digital de Saúde e Segurança do Trabalho (OBSERVATÓRIO, 2017).

O Observatório foi criado pelo SmartLab (laboratório multidisciplinar de gestão do conhecimento) e tem como objetivo mostrar estatísticas de acidentes de trabalho em todo o Brasil, como número de CATs registradas, valores gastos com benefícios acidentários,

dias perdidos de trabalho, mortes acidentárias, entre outros ([OBSERVATÓRIO, 2017](#)). Os dados podem ser visualizados por meio de números, gráficos, localizações geográficas e também possui diferentes tipos de filtros que podem ser utilizados. Apesar do Observatório contribuir para a visualização e sumarização dos dados dos acidentes de trabalho no Brasil, ele não utiliza técnicas de mineração de dados e nem de inteligência artificial (IA), as quais poderiam contribuir para automatização de tarefas, tomada de decisão e controle de processos.

A vasta quantidade de dados do Observatório, os quais podem ser importantes para nortear tomada de decisões, traz o desafio de converter tais dados em informações claras para políticas de controle e prevenção de acidentes e doenças ocupacionais. Analisar estes dados manualmente é inviável, devido ao grande volume da base de dados e por novos dados serem gerados frequentemente.

Contudo, na atualidade existem várias ferramentas computacionais que auxiliam e facilitam a realização de consultas e análises de dados mais complexas e, quando associadas a tecnologias de extração de conhecimentos, torna-se mais fácil destacar informações relevantes, bem como encontrar informações escondidas, ou seja, que dificilmente seriam encontradas com métodos tradicionais.

Uma dessas tecnologias de extração de conhecimentos é a mineração de dados. Ela é conhecida como a ciência de extrair conhecimentos úteis de grandes repositórios de dados a partir da aplicação de técnicas da estatística, inteligência artificial, aprendizado de máquina, recuperação da informação, dentre outros ([HAND, 2007](#)). Uma das etapas importantes para a extração de informação de maneira eficiente é o pré-processamento de dados.

O pré-processamento envolve conhecer detalhadamente a base de dados, detectando a qualidade dos dados, padronização, tipos de variáveis, transformações, tamanho da base e formas que possam colaborar para a eficiência da mineração e que se adequam à tarefa que será utilizada ([SCHMITT et al., 2005](#)).

A MD possui algumas tarefas como: classificação, agrupamento, regressão e associação. A escolha de qual tarefa deve ser utilizada depende do objetivo do processo. Ainda que diferentes tarefas da MD possam ser aplicadas na base de dados do MPT, uma das primeiras tarefas a serem exploradas é a busca por grupos que representem benefícios que possuam alguma similaridade entre si. Esses grupos de benefícios podem ser melhor investigados a fim de se identificar o que esses benefícios possuem em comum e porque eles foram colocadas no mesmo grupo.

No entanto para o MPT essa é uma importante iniciativa que pode contribuir para auxiliar os profissionais da área da saúde e segurança do trabalho à direcionar pesquisas e procedimentos visando a prevenção de acidentes e doenças ocupacionais e um melhor controle sobre os gastos com benefícios acidentários, em especial a tarefa de agrupamento de dados pode ajudar a traçar um perfil dos trabalhadores que necessitam de mais aten-

ção quanto aos riscos de sofrerem acidentes do trabalho, uma vez que esta tarefa separa objetos em grupos de acordo com suas características.

Outras iniciativas já foram feitas no sentido de estudar dados de acidentes de trabalho, mas foram feitas usando apenas uma localidade (BARTOLOMEU et al., 2002) ou então utilizando apenas um tipo de agrupamento (PIGNATA, 2016).

1.2 Objetivos

1.2.1 Objetivo geral

Analisar o desempenho de algumas técnicas de pré-processamento e agrupamento de dados, como *k-means*, *canopy* e *expectation maximization* (EM), os quais são algoritmos simples e eficientes, aplicados nos dados de benefícios previdenciários do Observatório Digital de Saúde e Segurança do Trabalho usando medidas para validação de agrupamento.

1.2.2 Objetivos específicos

- Produzir uma versão pré-processada da base que possa ser usada por diferentes técnicas de mineração de dados, de forma a eliminar atributos irrelevantes e normalizar os dados.
- Investigar e aplicar diferentes formas de converter os atributos categóricos para numéricos, em especial, considerando os atributos que possuem um número muito grande de possíveis valores, com o apoio do especialista de domínio.
- Criar diferentes versões da base de dados contendo registro de apenas um problema específico a fim de se ter uma visão mais detalhada do problema, como por exemplo, montar uma base considerando somente os benefícios decorrentes de acidentes.
- Analisar os diferentes modelos produzidos pelas técnicas de agrupamento de dados usando medidas de validação de agrupamento, a fim de identificar as melhores técnicas de agrupamento para o problema em questão.

2 Revisão Bibliográfica

2.1 Introdução

Este capítulo apresenta os conceitos teóricos necessários sobre acidentes de trabalho, benefícios previdenciários e mineração de dados para o melhor entendimento deste estudo.

A seção 2.2 fornece uma visão geral sobre acidente de trabalho e CAT. A seção 2.3 apresenta os tipos de benefícios prestados para pessoas que sofrem acidentes ou doenças de trabalho. A seção 2.4 discute como o Ministério Público do Trabalho gerencia os acidentes. A seção 2.5 explica o processo de KDD e cada etapa do mesmo. A seção 2.6 detalha a etapa de pré-processamento dentro do processo de KDD, assim como algumas técnicas comumente usadas. A seção 2.7 traz o conceito de mineração de dados também dentro do processo de KDD, mostrando suas tarefas, em especial a de agrupamento, pois é a tarefa que será utilizado neste estudo. A seção 2.8 apresenta diferentes medidas para validação de agrupamento. A seção 2.9 cita os trabalhos relacionados ao tema e suas colaborações para o Ministério Público. Por último, a seção 2.10 traz as considerações finais.

2.2 Acidente de Trabalho

Acidente de Trabalho é o acidente que acontece pelo exercício da ocupação a serviço da empresa, causando ferimentos corporais ou distúrbios funcionais, provocando a perda ou a diminuição da disposição para o trabalho, seja definitivo ou temporário, e em alguns casos podendo até levar a morte (SINCOVAGA, 2012).

Consideram-se também como acidente do trabalho aquele ocorrido durante o trajeto do empregado para casa ou para o serviço e as diversas doenças originadas pelo mesmo, como as decorrentes da repetição dos mesmos movimentos (LER - lesão por esforço repetitivo; DORT - distúrbios osteomusculares relacionados ao trabalho), aquelas originadas por muito esforço físico e mental e as obtidas por agentes tóxicos que afetam a saúde. Isto esclarece a obrigação da empresa solicitar exames admissionais, periódicos e demissionais (MPT, 2014).

Existem diversos motivos que ocasionam os acidentes de trabalho, os principais são: a falta de instrução para o trabalhador quanto ao uso de EPI (equipamento de proteção individual), atitudes precipitadas em ambientes perigosos, ausência ou desatenção na fiscalização do ambiente de trabalho, equipamentos antigos que não são substituídos e não passam por manutenção, e o não cumprimento de leis trabalhistas por parte da empresa (ADMMC, 2016).

No decorrer do ano de 2016, cerca de 2.265 pessoas vieram a óbito e foram registrados por volta de 578.900 acidentes de trabalho, sendo que os acidentes típicos consistem em 74,59% do total, os de trajeto 22,78% e as doenças do trabalho 2,63% (BRASIL, 2016). As consequências desses acidentes causam um grande impacto em toda a sociedade, direta e indiretamente, no âmbito empresarial, familiar, social e econômico (ALVARES; COSTA, 2015).

A Previdência Social gastou 33.7 milhões em benefícios no mês de dezembro de 2016, sendo 29.2 milhões em benefícios previdenciários e acidentários e, os demais, assistenciais (PREVIDÊNCIA, 2017). Portanto, investir na saúde do trabalhador é de grande importância para diminuir os acidentes de trabalho, ajudar na redução dos encargos previdenciários, preservar a imagem da empresa, aumentar a produtividade e assegurar uma melhor qualidade de vida para o empregado (ALVARES; COSTA, 2015).

2.2.1 Comunicação de Acidente de Trabalho - CAT

Quando ocorre um acidente de trabalho, a empresa deve informar à Previdência Social, mesmo que não houver afastamento, e isso é feito por meio da Comunicação de Acidente de Trabalho (CAT) (INSS, 2017).

A CAT é um documento que contém informações importantes a serem preenchidas sobre o empregado acidentado, seu empregador, o acidente ocorrido, as testemunhas, a lesão, o atendimento médico recebido e o diagnóstico (BARTOLOMEU et al., 2002). O seu principal objetivo é garantir a assistência acidentária ou até mesmo uma aposentadoria por invalidez ao trabalhador junto ao INSS, além de ser utilizada para fins de controle estatísticos e epidemiológicos junto aos órgãos federais (PANTALEÃO, 2016).

Existem três tipos de CAT: inicial, reabertura e óbito. O tipo inicial é quando acontece a primeira comunicação do acidente ou doença à Previdência, a de reabertura é quando o tratamento ou afastamento do empregado inicia novamente devido uma piora na lesão do acidente ou doença, e por último, a de óbito, que ocorre quando o empregado falece, sendo que sua abertura deve ocorrer imediatamente após a CAT inicial (CIPA, 2011).

A comunicação deve ser realizada em até 24 horas úteis, em quatro vias, as quais deverão ser destinadas ao INSS, ao segurado ou dependente, ao sindicato de classe do trabalhador e à empresa. Caso a empresa se recuse a abrir uma CAT, o próprio empregado, o dependente do empregado, a entidade sindical, o médico ou a autoridade pública poderão abrir a mesma e a empresa ficará sujeita à multa (INSS, 2017).

Durante o período de 2014 à 2016 houve uma média de 515.590 acidentes de trabalho com CAT registrada, sendo 390.061 de acidentes típicos, 110.367 de trajeto e 15.162 de doença do trabalho, além de ter uma média de 2.543 óbitos registrados (BRASIL, 2016). O preenchimento correto e completo da CAT é muito importante para se obter dados estatísticos confiáveis, que possam servir como proposição de políticas preventivas e

a estudos voltados para o conhecimento da real extensão dos acidentes do trabalho no Brasil ([MORAES, 2012](#)).

2.3 Benefícios Previdenciários

Benefícios são prestações monetárias pagas pela Previdência Social aos segurados (pessoas cobertas pelo sistema previdenciário) ou aos seus dependentes de forma a atender a cobertura das ocorrências de doença, idade avançada, invalidez, morte, maternidade, salário-família, auxílio reclusão e pensão por morte do segurado ([BRASIL, 2016](#)).

Os benefícios são classificados em espécie. Essa classificação foi criado pelo INSS para esclarecer particularidades de cada tipo de benefício monetário existente ([PJERJ, 2014](#)). Há diversas espécies, porém, como o tema tratado é acidentes de trabalho, serão descritos algumas espécies decorrentes de incapacidade para o trabalho.

2.3.1 Auxílio-doença acidentário

O auxílio-doença é um benefício pago ao segurado do INSS que comprovar, por meio de perícia médica, que está temporariamente incapaz de exercer suas atividades do trabalho em consequência de doença ou acidente ocorrido no exercício do trabalho. O código da espécie desse auxílio é o 91. Ele é isento de carência (número mínimo de meses pagos ao INSS para que o cidadão, ou em alguns casos o seu dependente, possa ter direito de receber um benefício) e corresponde a alíquota de 91% sobre o salário de benefício (consiste na média aritmética simples dos maiores salários de contribuição correspondentes a oitenta por cento do período contributivo decorrido desde a competência de julho de 1994 até a data do início do benefício) ([COSTA, 2017](#)).

2.3.2 Aposentadoria por invalidez acidentária

A aposentadoria por invalidez é um benefício pago ao segurado que é considerado incapaz para o trabalho e de se reabilitar para exercer o exercício de atividade que lhe garanta o sustento, decorrente de acidente ou doença do trabalho. Caso o aposentado voltar à atividade, a aposentadoria é cancelada. O código da espécie desse auxílio é o 92. Ele é isento de carência e tem a alíquota de 100% sobre o salário de benefício ([BRASIL, 2016](#)).

2.3.3 Pensão por morte por acidente de trabalho

Este é um benefício concedido aos dependentes do segurado que falece decorrente a acidente de trabalho e tem como objetivo suprir as necessidades dos mesmos. O código da espécie dessa auxílio é o 93. Aqui a carência não é mais exigida, porém, a pessoa que

falecer deve possuir a qualidade de segurado para que seus dependentes tenham direito a este benefício. Contém uma alíquota de 100% sobre o valor da aposentadoria que o segurado recebia (COSTA, 2017).

2.3.4 Auxílio-acidente

O auxílio-acidente é um benefício concedido, como indenização, ao segurado empregado, trabalhador avulso e segurado especial que ficaram com sequelas definitivas após sofrerem lesões decorrentes de acidente de qualquer natureza. O código da espécie dessa auxílio é o 94. Ele independe de carência e possui alíquota de 50% sobre o salário de benefício (ARAÚJO, 2011).

2.4 Políticas do Ministério Público do Trabalho para gerenciar acidentes

O Ministério Público do Trabalho (MPT) é a área do Ministério Público da União (MPU) que tem como objetivo controlar a execução da legislação trabalhista. Compete ao MPT proporcionar a ação civil pública no campo da Justiça do Trabalho para resistência de interesses coletivos, quando desacatados direitos sociais constitucionais assegurados aos trabalhadores (MPT, 2015).

O MPT tem procurado preservar o ambiente de trabalho o mais sadio e seguro possível, aderindo todas as soluções necessárias para distanciar ou minimizar os riscos à saúde e à plenitude física dos trabalhadores. Porém, ele necessita do cumprimento, tanto por parte do empregado quanto do empregador, das normas que regulam este quesito (MPT, 2013).

Dentre as normas que regulam a preservação da saúde e da segurança no meio ambiente do trabalho estão: o fornecimento, instrução e uso efetivo dos Equipamentos de Proteção Individual (EPI), constituição e funcionamento das Comissões Internas de Prevenção de Acidentes (CIPA) e dos Serviços Especializados em Engenharia de Segurança e em Medicina do Trabalho (SESMT), existência e implementação de Programa de Controle Médico de Saúde Ocupacional (PCMSO) e Programa de Prevenção de Riscos Ambientais (PPRA), pagamento de adicional para trabalho perigoso ou insalubre, existência de local apropriado para as refeições dos trabalhadores, instalações sanitárias adequadas separadas por sexo, dotados de chuveiros, lavatórios, vestiários e armários individuais, fornecimento de água potável e de copos descartáveis, a existência de extintores de incêndio portáteis para combate inicial de fogo, entre outras normas (VALADA, 2015).

Em 2017, o MPT juntamente com a OIT (Organização Internacional do Trabalho), lançaram o Observatório Digital de Saúde e Segurança do Trabalho. Este observatório é uma ferramenta online que tem como objetivo tornar público e acessível as informações

sobre a saúde e segurança do trabalhador que antes se encontravam perdidas em banco de dados governamentais, além de possuir grande capacidade para contribuir com o andamento, acompanhamento e a avaliação de projetos, políticas públicas e programas de prevenção de acidentes e doenças no trabalho (OIT, 2017).

Na Figura 1 é representado a tela inicial do observatório, a qual já de imediato, inicia na aba de Frequência do menu e apresenta diversas informações para o público, como: o total de gastos da Previdência com benefícios acidentários, quantidade de dias de trabalho perdidos com afastamentos, quantidade de acidente desde 2012 até a atualidade, quantidade de mortes acidentárias notificadas e geolocalização por municípios. Possui também diversos filtros por CAT's ou por afastamentos pelo INSS, como: parte do corpo atingida, agente causador, natureza da lesão, tipo de acidente registrado, tipo e causa do afastamento, categoria CID, Classe da Atividade Econômica (CNAE) e sexo.

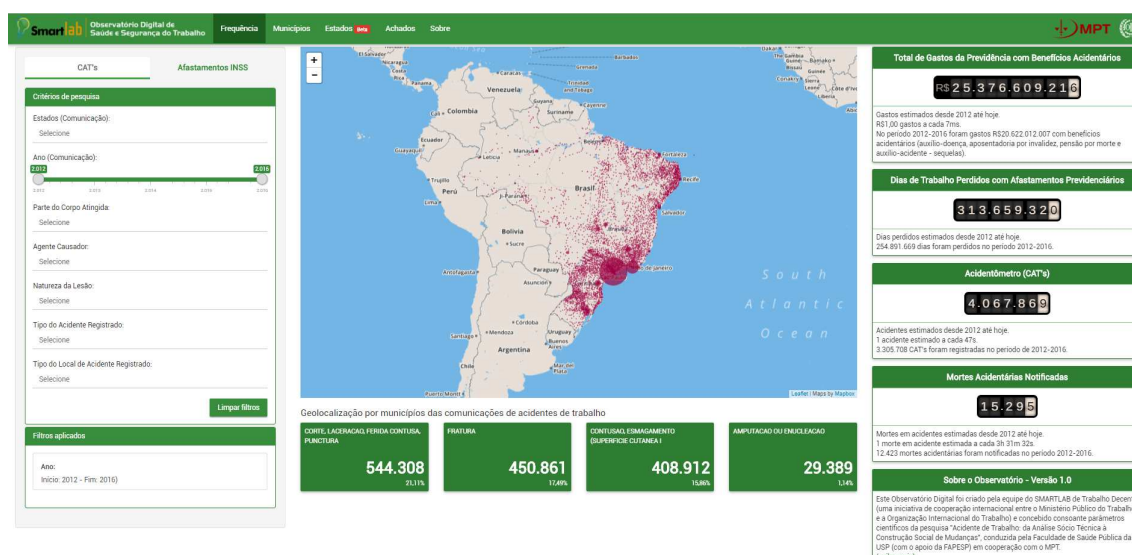


Figura 1 – Tela inicial do Observatório Digital de Saúde e Segurança do Trabalho

Fonte – (OBSERVATÓRIO, 2017)

A segunda, terceira e quarta abas do menu do observatório (Municípios, Estados, Achados) contém diferentes tipos de gráficos, tanto de acidentes de trabalho quanto de afastamentos previdenciários acidentários (Figura 2), que podem ser filtrados de diversas maneiras (município, estados, distribuição geográfica, vítimas menores de 18 anos, partes do corpo mais frequentemente atingidas, dentre outros), facilitando a visualização e compreensão das estatísticas apresentadas.

A última aba do menu do observatório (Sobre) contém informações sobre os colaboradores, as novidades das próximas versões, as tecnologias e bancos de dados utilizadas para a criação do observatório, descrições das principais funcionalidades e possui também os conjuntos de dados o qual serão analisados neste trabalho.

Apesar de todas as informações que podem ser extraídas do observatório, o MPT necessita mais do que apenas visualizações e sumarizações dos dados, necessita saber qual



Figura 2 – Gráficos do Observatório Digital de Saúde e Segurança do Trabalho

Fonte – (OBSERVATÓRIO, 2017)

o perfil dos trabalhadores que mais sofrem acidentes, quais os ramos empresariais oferecem maiores riscos a saúde do trabalhador e até mesmo estimar o quanto será gasto com benefícios no ano seguinte, a partir da base de dados dos anos anteriores, recursos que o observatório atualmente não dispõe.

2.5 Descoberta de conhecimento em base de dados

A descoberta de conhecimento em base de dados, ou *Knowledge Discovery in Databases* (KDD) é o processo global de transformação de dados em informações. Devido a isto, diversas áreas de conhecimentos estão relacionadas neste processo, como: matemática, estatística, banco de dados, inteligência artificial, reconhecimento de padrões e visualização de dados (FAYYAD et al., 1996).

O KDD é reconhecido como um processo constituído por várias etapas operacionais. A complexidade deste processo está em entender e decifrar devidamente os fatos observáveis e em associar dinamicamente tais interpretações de forma a decidir quais ações devem ser executadas em cada caso. Cabe ao analista humano a difícil tarefa de guiar a execução do processo de KDD (GOLDSCHMIDT; PASSOS, 2005). Na Figura 3 é mostrado as cinco etapas que compõem o processo de KDD.

2.5.1 Seleção dos dados

Esta etapa envolve identificar quais informações nas bases de dados existentes, devem ser realmente consideradas durante o processo de KDD (GOLDSCHMIDT; PASSOS, 2005). A seleção dos dados pode ter dois aspectos diferentes: de atributos ou de

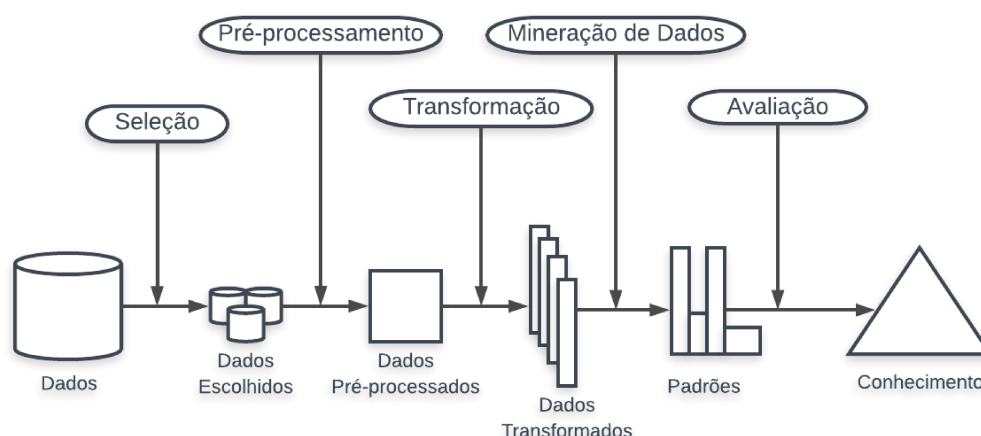


Figura 3 – Etapas do processo de KDD

registros, que devem ser submetidos ao processo. Por exemplo, o nome do cliente é um atributo desnecessário quando se trata de um modelo de prever comportamento de novos clientes quanto ao pagamento de futuros créditos a eles concedidos. No entanto, o atributo data de nascimento é essencial em um modelo para estimar o apólice de seguro de vida do cliente. Já no contexto da seleção de registros, pode ser que seja inviável usar todo o conjunto de dados por causa da ausência de algum atributo ou por algum outro motivo, então um subconjunto dos registros terá que compor o conjunto de dados a ser analisado (GOLDSCHMIDT; PASSOS, 2017).

2.5.2 Pré-Processamento dos dados

O objetivo da etapa de pré-processamento é modificar os dados de entrada para um formato mais apropriado para análises posteriores. Os passos envolvidos no pré-processamento de dados abrangem a limpeza dos mesmos para remoção de ruídos, a fusão de dados de múltiplas fontes, observações duplicadas e características que sejam importantes à tarefa de mineração de dados (TAN; STEINBACH; KUMAR, 2009).

2.5.3 Transformação dos dados

A etapa de transformação dos dados requer a combinação de diferentes base de dados alterando seu formato e evoluindo as informações. É necessário levar em consideração os tipos de algoritmos que serão utilizados a fim de que os dados sejam preparados, por exemplo, alguns algoritmos trabalham somente com valores numéricos, já outros só trabalham com valores categóricos. Diversas técnicas podem ser utilizadas nesta etapa, como: normalização (ajusta escalas de valores dos atributos para um mesmo intervalo), generalização (converte valores específicos para genéricos) e suavização (remove valores incorretos) (FABRICIO, 2017).

2.5.4 Mineração de Dados

A Mineração de Dados é a principal etapa do processo de KDD, sendo responsável pela escolha dos melhores algoritmos a serem utilizados no problema em questão e efetuado a busca efetiva por informações úteis (GOLDSCHMIDT; PASSOS, 2005). Dentre as tarefas de mineração de dados, pode-se destacar classificação, regressão, associação e agrupamento.

2.5.5 Avaliação dos resultados

A última etapa do processo de KDD consiste em interpretar e empregar os conhecimentos adquiridos na tomada de decisão. São apresentadas nesta fase também as medidas de desempenho (REZENDE, 2003).

Os resultados do processo de KDD podem ser interpretados e visualizados de diferentes formas, como gráficos, tabelas, diagramas, relatórios demonstrativos, entre outros. É importante o envolvimento de todos os participantes nesta fase, para avaliarem de forma sensata os resultados (CASTANHEIRA, 2008).

As próximas seções vão dar destaque a tarefa de pré-processamento e agrupamento, pois estas estão relacionadas aos objetivos do projeto e são as etapas que consomem mais tempo no processo de KDD (MANNILA, 1996).

2.6 Pré-processamento

Nesta seção será mostrado com mais detalhes a etapa do Pré-processamento de dados dentro do processo de KDD. Esta etapa tem a função de aprimorar a qualidade dos dados fazendo com que os processos de MD fiquem mais eficientes (HAN; PEI; KAMBER, 2011).

As subseções 2.6.1 a 2.6.3 descrevem alguns elementos que definem e comprometem a qualidade dos dados e também descrevem sobre as técnicas de pré-processamento voltadas para instâncias, já as subseções 2.6.4 à 2.6.7 destacam as técnicas voltadas para atributos.

2.6.1 Qualidade dos dados

Três elementos definem qualidade dos dados: precisão, completude e consistência (HAN; PEI; KAMBER, 2011). Porém, não se pode esperar que uma base de dados tenha dados perfeitos. Ela pode conter problemas pertinentes a erro humano, falhas na coleta de dados ou limitações nos dispositivos de medição. A maioria das vezes a MD é aplicada em dados que foram coletados para um outro propósito, por isso eles podem estar armazenados em diferentes formatos ou conter inconsistências (TAN; STEINBACH; KUMAR, 2009). As inconsistências mais comuns são:

- Ruídos: são erros aleatórios que podem implicar na distorção de um valor ou a adição de objetos falsos (TAN; STEINBACH; KUMAR, 2009).
- *Outliers*: são objetos de dados com características distintas dos demais objetos do mesmo conjunto de dados ou são valores anormais de um atributo (BARNETT; LEWIS, 1974).
- Valores Ausentes: podem ocorrer por vários motivos. Por exemplo, em caso de entrevista, o entrevistado pode ter receio de informar idade ou renda, em formulários, alguns campos não são obrigatórios preencherem (WITTEN; FRANK; HALL, 2011). Independente da ocasião, os valores ausentes devem ser levados em consideração durante a análise de dados (TAN; STEINBACH; KUMAR, 2009). Existem diferentes estratégias para lidar com dados ausentes, como:
 - Eliminar linha: essa estratégia é simples porém não é a mais indicada, pois pode resultar em um baixo desempenho se a quantidade de linhas com valores ausentes for alta. É mais utilizado quando a linha contém vários atributos com valores ausentes (HAN; PEI; KAMBER, 2011).
 - Ignorar Valores Ausentes durante a Análise: Muitas abordagens de MD podem ser alteradas para ignorar valores ausentes. Considere que objetos estejam sendo agrupados e a equivalência de pares de objetos tenha que ser calculada. Caso um ou os dois objetos de um par possuir valores ausentes, então a equivalência pode ser calculada pelos atributos não ausentes (TAN; STEINBACH; KUMAR, 2009).
 - Imputar Valores Ausentes: os valores ausentes podem ser substituídos de acordo com vários critérios. No caso de variáveis numéricas, estes valores podem ser substituídos pela média do campo, já para as variáveis categóricas, podem ser substituídas pela moda. Um outro modo seria substituir por alguma constante determinada pelo analista (LAROSE, 2005).
- Dados Duplicados: um conjunto de dados pode conter objetos de dados que estão duplicados. Deve-se ter cuidado para evitar combinar inconscientemente objetos de dados que sejam similares e não duplicados, como duas pessoas diferentes com nomes iguais. O processo de lidar com instâncias duplicadas que detecta e corrige este problema é chamado de deduplicação (TAN; STEINBACH; KUMAR, 2009).

A baixa qualidade dos dados leva a baixa qualidade dos resultados da MD. Portanto, o pré-processamento constitui diferentes técnicas que podem ser usadas para melhorar a MD em relação ao tempo, custo e qualidade (HAN; PEI; KAMBER, 2011). Nas próximas subseções serão apresentadas as técnicas mais importantes.

2.6.2 Agregação

A Agregação é a união de duas ou mais transações em uma única transação. Isto reduz o tempo de processamento e necessita de menos memória, possibilitando uso de algoritmos de MD mais complexos. Atributos quantitativos são normalmente agregados atribuindo uma soma dos valores ou a média, já atributos qualitativos podem ser omitidos ou resumido como o conjunto de todos atributos daquele ambiente. Uma desvantagem desta técnica é a perda de detalhes (TAN; STEINBACH; KUMAR, 2009).

2.6.3 Amostragem

Esta técnica é usada para escolher subconjuntos dos objetos de dados a serem examinados. A amostragem é interessante pois reduz o custo e o tempo para processar os dados em relação ao conjunto de dados completo, desde que a mesma seja representativa. Uma amostra é representativa se tiver praticamente a mesma propriedade (de interesse) da base de dados original (BAOHUA; FEIFANG; HUAN, 2000).

2.6.4 Redução de dimensionalidade

Conjuntos de dados podem possuir um grande número de atributos. Com a técnica de redução de dimensionalidade pode ser excluído características desnecessárias e diminuir o ruído, além de levar a uma forma mais clara, facilitando a visualização dos dados. Porém, tudo isso acontece em partes, pois existe a maldição da dimensionalidade, que é quando ocorre o aumento da dimensionalidade dos dados e as análises se tornam cada vez mais difíceis, fazendo com que os dados se espalham cada vez mais no espaço que eles se encontram (TAN; STEINBACH; KUMAR, 2009).

2.6.5 Seleção de subconjuntos de atributos

Utilizar apenas um subconjunto de atributos é outra maneira de reduzir a dimensionalidade, porém, um conjunto de dados pode conter atributos redundantes ou irrelevantes. Os atributos redundantes duplicam as informações apresentadas em um ou mais atributos, já os irrelevantes não possuem informações importantes para a tarefa prevista. Mesmo que alguns atributos irrelevantes e redundantes possam ser excluídos utilizando bom senso ou conhecimento do domínio, escolher o melhor subconjuntos de atributos exige uma abordagem sistemática (TAN; STEINBACH; KUMAR, 2009). Em geral, três abordagens são usadas para a seleção de atributos, são eles (KOHAVI; JOHN, 1997):

- *Embedded*: os atributos são selecionados naturalmente como parte do algoritmo de MD.

- Filtros: os atributos são selecionados utilizando uma abordagem que independe da MD e antes da execução da mesma.
- *Wrappers*: utilizam o algoritmo de MD para encontrar o melhor subconjunto de atributos.

2.6.6 Criação de atributos

A partir dos atributos originais é possível criar um novo conjunto que contenha as informações importantes e um número menor de atributos, possibilitando obter os benefícios de redução de dimensionalidade. Existem três métodos comumente usados para criar novos atributos, que são (TAN; STEINBACH; KUMAR, 2009):

- Extração de Características: cria um novo conjunto de atributos a partir da base original. Este método é mais utilizado em processamento de imagens.
- Mapeamento de dados com um Novo Espaço: cria um novo conjunto de atributos através da Transformada de Fourier.
- Construção de Recursos (Características): os atributos originais podem ter todas as informações necessárias, porém, podem estar em uma forma inadequada para o algoritmo de MD. Portanto, este método cria um novo conjunto de atributos com a forma adequada para o algoritmo de MD.

2.6.7 Discretização e Binarização

A discretização é a técnica que converte um atributo contínuo em categórico. Para realizar esta conversão é preciso estabelecer o número de categorias a serem usadas e definir como estruturar os valores contínuos deste atributo para essas categorias (MENDES, 2011). O resultado da discretização de atributos contínuos pode ser apresentado como conjunto de intervalos $[x_0, x_1]$, $[x_1, x_2]$, ... $[x_{n-1}, x_n]$ onde x_0 e x_n podem ser $+\infty$ ou $-\infty$, respectivamente ou, em um formato semelhante, como uma série de desigualdades $x_0 < x \leq x_1, \dots, x_{n-1} < x \leq x_n$ (MENDES, 2014).

Por outro lado, a binarização é a técnica que converte atributos contínuos ou discretos em binários. Um exemplo simples de binarização pode ser feito com um atributo que define o campo sexo de uma pessoa. Este campo pode conter o valor F para o sexo feminino ou M para o sexo masculino. Convertendo para binário ficaria $F = 1$ e $M = 0$ (MENDES, 2011). A seção 3.3 apresenta dois exemplos de binarização.

2.7 Mineração de Dados

A mineração de dados (MD) consiste no processo de exploração automática de informações úteis em grandes repositórios de dados. As técnicas de MD agem em grandes bancos de dados com o objetivo de descobrir padrões úteis, os quais por outros tipos de análises poderiam não serem detectados (TAN; STEINBACH; KUMAR, 2009).

Dentro do processo de KDD, a etapa de MD requer a escolha da técnica e algoritmo que serão utilizados na tarefa proposta. Após ter feito a escolha, será necessário desenvolver o algoritmo, adaptando-o ao problema proposto e então executá-lo para obter-se os resultados que serão analisados na fase de interpretação e avaliação do resultado (CASTANHEIRA, 2008).

A MD possui diversas tarefas como: classificação, regressão, associação e agrupamento. Essas tarefas podem apresentar diferentes tipos de conhecimentos conforme será apresentado nas subseções 2.7.1 à 2.7.4, porém neste trabalho será dada uma maior atenção à tarefa de agrupamento, que será usada na mineração dos dados do MPT.

2.7.1 Classificação

A Classificação é a tarefa de aprender uma função f que mapeie um conjunto de atributos x em um conjunto de variáveis predefinidas y , denominadas rótulos de classes. A função f também é conhecida informalmente como modelo de classificação. Essa modelagem pode ser descritiva, ou seja, pode servir como ferramenta explicativa para se distinguir entre objetos e classes diferentes, ou pode ser preditiva, que prevê o rótulo de classe de registros não conhecidos. Técnicas de classificação são mais apropriadas para prever ou descrever conjuntos de dados com categorias nominais (não possuem uma ordem definida) ou binária, sendo menos efetiva para categorias ordinais (que possuem uma ordem definida) (TAN; STEINBACH; KUMAR, 2009).

2.7.2 Regressão

Enquanto a classificação prevê valores categóricos, a regressão é aplicada a valores numéricos, tendo como propósito prever dados históricos existentes em uma base de dados, isto é, assimila a busca por uma função que esquematize os registros de um banco de dados para valores reais. Outras áreas como Redes Neurais e Estatística apresentam ferramentas para implementação da tarefa de regressão (MICHIE; SPIEGELHALTER; TAYLOR, 1994).

2.7.3 Regras de Associação

Esta tarefa é utilizada para encontrar padrões que relatem qualidades profundamente associadas nos dados. Os padrões encontrados são geralmente exibidos na forma

de subconjuntos de características ou regras de implicação. A análise de associação tem como objetivo extrair padrões relevantes de forma ágil, isso acontece devido ao tamanho exponencial de sua área de busca. O reconhecimento de páginas Web acessadas simultaneamente e a descoberta de genes que possuam funcionalidade associada são exemplos de aplicações de análise de associação (TAN; STEINBACH; KUMAR, 2009).

2.7.4 Agrupamento

Métodos de agrupamento ou *clustering* são utilizados para dividir objetos de dados em grupos, ou então, como um passo de pré-processamento para outros algoritmos (TAN; STEINBACH; KUMAR, 2009). São denominados como aprendizado não supervisionado devido as informações do rótulo de classe não estar presente (HAN; PEI; KAMBER, 2011).

A tarefa de agrupamento busca dividir o conjunto de dados em grupos homogêneos, ou seja, maximiza a similaridade dos dados dentro do grupo e minimiza os que permanecem fora do mesmo. Um grupo é um conjunto de dados que são similares entre si e diferente dos dados de outros grupos (LAROSE, 2005).

De acordo com Amo e Roc (2003), os métodos de agrupamento podem ser classificados nas seguintes categorias: hierárquicos, particionais, baseados em densidade e uma grande quantidade de outros métodos que utilizam diferentes técnicas (AMO; ROC, 2003). Nas subseções a seguir, serão descritas as características das principais categorias, assim como os algoritmos mais populares no processo de agrupamento.

2.7.4.1 Métodos Hierárquicos

Os métodos hierárquicos constituem um conjunto de dados em uma estrutura hierárquica conforme a proximidade dos elementos. Normalmente, os grupos são representados por um dendograma, que é uma árvore que divide a base de dados em subconjuntos menores. Neste dendograma, um elemento é representado pela folha e o agrupamento de todos os elementos é representado pela raiz. É necessário definir uma distância de corte para apresentar quais serão os grupos formados, portanto, é fundamental ter um conhecimento sobre a estrutura dos dados e do objetivo da análise. O dendograma pode ser criado de duas formas: aglomerativa ou divisiva (DONI, 2004).

Na forma aglomerativa, é iniciado das folhas para a raiz (*bottom-up*). Cada elemento é considerado um grupo, obtendo-se assim n grupos. A cada etapa é calculado a distância entre cada par de grupo e salvo em uma matriz de dissimilaridade simétrica (matriz em que cada registro representa a distância entre pares de elementos). Feito isso, junta-se dois grupos com distâncias mínimas e atualiza a matriz. Este procedimento continua até que todos os elementos se encontrem em um único grupo ou até que se tenha um ponto de parada (CASSIANO, 2014).

A forma divisiva é praticamente o inverso da aglomerativa, é iniciado da raiz para

as folhas (*top-down*). Todos os elementos começam em um único grupo, e a cada etapa um grupo é dividido em dois grupos menores. Este procedimento continua até que se obtenha n grupos ou até que se tenha um ponto de parada (VALE, 2005).

Um exemplo de dendograma representando agrupamento hierárquico é apresentado na Figura 4, no qual a forma aglomerativa se desloca de baixo para cima e a forma divisiva de cima para baixo.

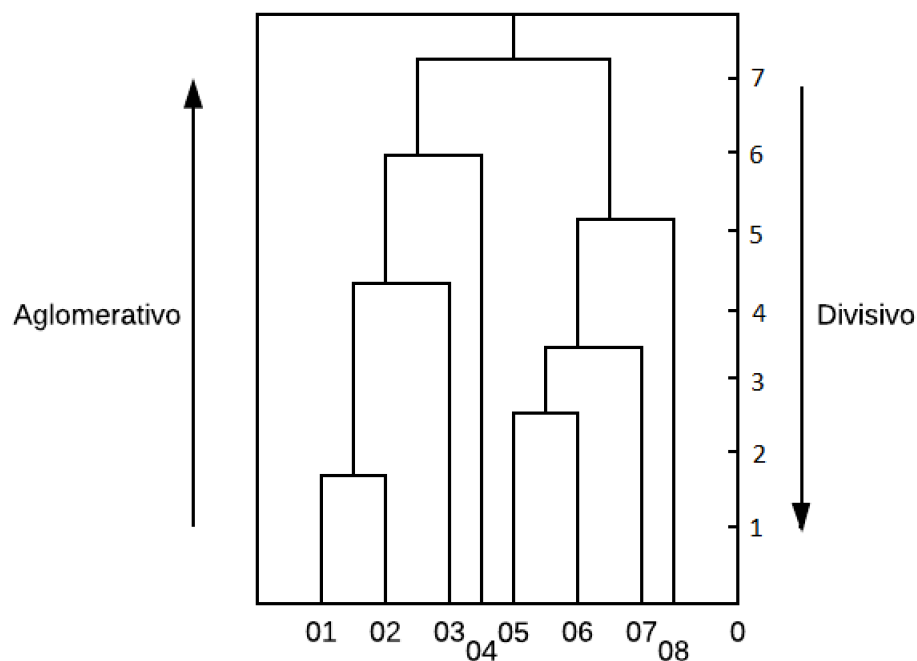


Figura 4 – Dendograma representando agrupamento hierárquico

2.7.4.2 Métodos Particionais

Os métodos particionais tem a função de agrupar registros em k grupos, onde k é a quantidade de grupos estabelecidos a priori. Este método é executado diversas vezes com valores diferentes de k , selecionando resultados que expressem uma melhor interpretação dos grupos ou melhor representação gráfica (DONI, 2004).

As vantagens desses métodos em comparação aos métodos hierárquicos são que eles podem trabalhar com base de dados maiores e são capazes de alterar um agrupamento com a evolução do algoritmo. Já as desvantagens é que são sensíveis às condições iniciais, sendo capazes de apresentar diferentes resultados a cada execução, e por terem que escolher um número de agrupamento a priori, sendo que, caso o número de agrupamentos não for o melhor, pode causar incorretas interpretações sobre a estrutura dos dados (VALE, 2005).

Um dos algoritmos particionais mais conhecido é o *k-means* (k-médias). Ele é muito utilizado para agrupamento por ser de fácil implementação, eficiente e simples (JAIN, 2010). Na subseção a seguir será detalhado o funcionamento deste algoritmo.

2.7.4.2.1 *K-means*

O *k-means* é um algoritmo que necessita que um número k de grupos que pretende obter-se seja estipulado. Ele funciona com o conceito de que cada grupo contém um centro, ou centroide, que é calculado com base nas características dos dados de cada grupo. A cada novo dado inserido ao grupo, o centroide do mesmo é recalculado baseado na média das características dos dados do grupo (GROSS, 2014).

O Algoritmo 1 mostra o funcionamento básico do *k-means*. Primeiramente é determinado pelo usuário pontos como centroides iniciais, o qual indica o número de grupos desejado. A seguir, são calculadas as distâncias de todos os elementos do conjunto de dados em relação aos k centroides. Cada elemento é adicionado ao grupo cuja distância ao seu centroide seja a menor. Após isso, os centroides de cada grupo são atualizados. O processo de atribuição de elementos a grupos e atualização dos centroides se repete até que nenhuma mudança ocorra, ou seja, os grupos se estabilizam (TAN; STEINBACH; KUMAR, 2009).

Algoritmo 1: Algoritmo *k-means* básico

início| Selecione k pontos como centroides iniciais;**repita**| | Forme k grupos atribuindo cada ponto ao seu centroide mais próximo;

| | Recalcule o centroide de cada grupo;

| **até que os centroides não mudem;****fim**

Fonte – (TAN; STEINBACH; KUMAR, 2009)

Na Figura 5 é ilustrado o processo do método *k-means* para $k = 3$. Os centroides são representados pelo símbolo “+”. Cada iteração representa o estado após recalculando o centroide de cada grupo, exceto a primeira que os centroides foram escolhidos aleatoriamente. Foram necessárias seis iterações para que os grupos estabilizassem.

Este método é prático e computacionalmente eficiente, porém tem suas desvantagens, é sensível a ruídos, *outliers* e não pode lidar com grupos de densidades diferentes (OLIVEIRA, 2016).

2.7.4.2.2 *Canopy*

O algoritmo *Canopy* é utilizado em grandes e multidimensionais bases de dados devido ao fato de ter um baixo custo computacional, pois requer apenas uma passagem sobre os dados e utiliza métricas de distância fáceis e rápidas de calcular. Sua função é agilizar a concatenação dos elementos em grandes bases, pois alguns algoritmos sofrem

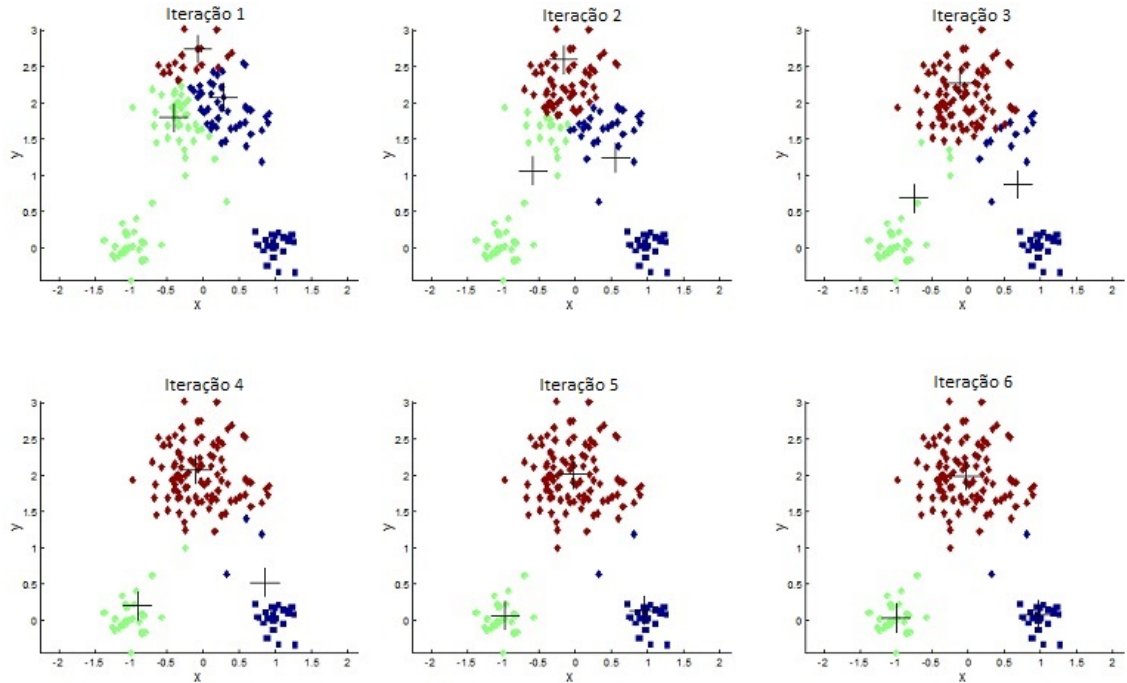


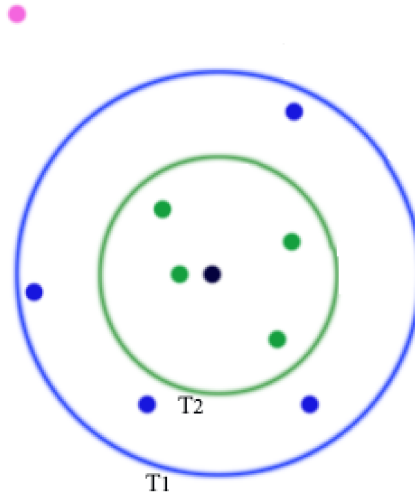
Figura 5 – Exemplo de execução do *k-means* em uma base com 3 grupos

Fonte – (PANDRE, 2009)

limitações pela alta quantidade de dados. Ele possui dois valores, T_1 (distância solta) e T_2 (distância apertada) e realiza os seguintes passos (MIRANDA, 2016):

1. Inicia com a base de dados a ser agrupada.
2. Seleciona e remove um elemento do conjunto de dados como centro de um novo *canopy*.
3. Para cada elemento deixado no conjunto, atribua-o ao novo *canopy* se a distância for menor que T_1 .
4. Se a distância do elemento é menor que T_2 , remova-o do conjunto original.
5. Repita o procedimento a partir do passo 2 até que não haja mais elementos no conjunto para ser agrupado.

Na Figura 6 é representado um agrupamento utilizando o algoritmo *canopy*. Para definir a região *canopy*, foi desenhado um círculo (azul) centralizado em um ponto de dados. Pontos fora deste círculo são considerados muito longe (ponto rosa). No entanto, se aplicar essa definição a todos os pontos, irá conter tantos *canopies* quanto o número de pontos, aumentando muito o processamento. Portanto, foi desenhado um círculo menor (verde) dentro do círculo maior, de forma que os pontos de dados dentro do pequeno círculo (pontos verdes) não possam formar seu próprio *canopy*. Cada *canopy* pode se

Figura 6 – Exemplo do agrupamento *canopy*

Fonte – (HUI, 2017)

sobrepor e o grau de sobreposição será afetado pela escolha de T_1 . A escolha de T_2 afeta quantos *canopies* serão formados (HO, 2011).

O *canopy* é muito utilizado como uma etapa inicial em técnicas de agrupamentos mais rigorosas como o *k-means* (MAHOUT, 2017).

2.7.4.2.3 Expectation Maximization (EM)

O algoritmo EM é um modelo estatístico que depende de parâmetros iniciais para iniciar o processo de iteração. Ele é similar ao procedimento do *k-means*, no qual o conjunto dos parâmetros é recomputado até que o valor de convergência desejado seja alcançado. A premissa básica do algoritmo é que deve encontrar um ponto ótimo a partir do qual não seja preciso realizar outras tentativas com novos parâmetros, de forma a representar os dados na forma mais verossímil possível (CAETANO; MACHADO, 2009).

O modelo mais utilizado do algoritmo EM é aquele cuja função de distribuição de probabilidades se dá por uma mistura de Gaussianas representada pela equação 2.1, onde x_i se refere a um objeto, $\mathcal{N}$ é uma Gaussiana, v_i é o centro da i -ésima Gaussiana, Σ_i é a matriz de covariância da i -ésima Gaussiana, π_i é uma probabilidade associada à i -ésima Gaussiana e k é o número de Gaussianas que compõem a mistura (CAMPELLO, 2014).

$$p(x_j) = \sum_{i=1}^k \pi_i \mathcal{N}(x_j | v_i, \Sigma_i) \quad (2.1)$$

Por ser um algoritmo de otimização, o EM visa maximizar a função de verossimilhança em 2 passos. O primeiro passo é o E (*Expectation*) que avalia as probabilidades a posteriori $\mu_{ij} (i = 1, \dots, k; j = 1, \dots, N)$ representado pela Equação 2.2. O segundo passo é

o M (*Maximization*), o qual ajusta o modelo visando maximizar a verossimilhança (CAMPELLO, 2014).

$$\mu_{ij} = \frac{\pi_i \mathcal{N}(x_j | v_i, \Sigma_i)}{\sum_{l=1}^k \pi_l \mathcal{N}(x_j | v_l, \Sigma_l)} \quad (2.2)$$

O algoritmo 2 representa todo o processo do EM.

Algoritmo 2: Algoritmo EM

início

Selecione um conjunto inicial de parâmetros de modelos;
(Assim como em *k-means*, isto pode ser feito aleatoriamente, em uma diversidade de formas.)

repita

Etapla da Expectativa Para cada objeto, calcule a probabilidade de que cada objeto pertença a cada distribuição, i.e., calcule $\text{prob}(\text{distribuição } j | x_i, \theta)$;

Etapla da Maximização Dadas as probabilidades da etapa da expectativa, encontre as novas expectativas dos parâmetros que maximizem a probabilidade esperada;

até que *Os parâmetros não mudem*;

(De forma alternativa, para se a mudança nos parâmetros estiver abaixo de um limite especificado.)

fim

Fonte – (TAN; STEINBACH; KUMAR, 2009)

2.7.4.3 Métodos baseados em densidade

Enquanto os métodos particionais definem apenas grupos de formato circular ou esférico, os métodos baseados em densidade são capazes de identificar grupos de formato irregular ou arbitrário, além de serem eficientes para encontrar ruídos (CASSIANO, 2014).

Para entender a ideia dos métodos baseados em densidade, ao observar a Figura 7, pode-se perceber facilmente, que no conjunto 1 possui grupos circulares, no 2 arbitrários e no 3 a presença de ruídos. A principal razão pela qual se tem este reconhecimento é que, dentro de cada grupo, existe uma densidade específica de pontos que é consideravelmente maior do que fora do mesmo. Além disso, a densidade dentro das áreas de ruído, é menor que a densidade em qualquer um dos grupos (ESTER et al., 1996). O cérebro humano reconhece os grupos e ruídos da Figura 7, utilizando automaticamente o conceito de grupos formados por densidade (CASSIANO, 2014).

Neste contexto, o algoritmo DBSCAN é um dos mais conhecidos e reflete conceitos importantes, que servem de apoio para qualquer abordagem baseada em densidade

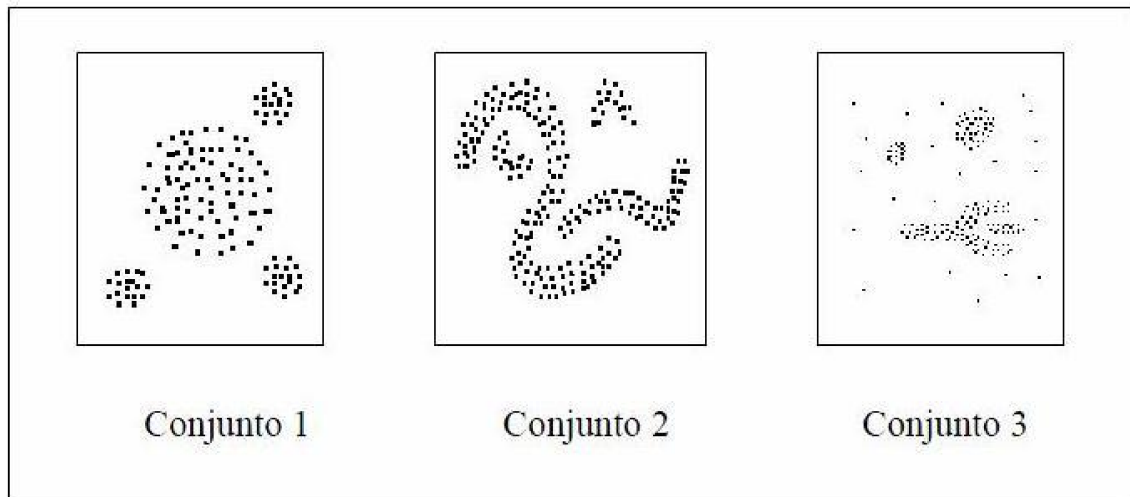


Figura 7 – Conjunto com grupos circulares (conjunto 1), não circulares (conjuntos 2 e 3) e com ruídos (conjunto 3).

Fonte – (ESTER et al., 1996)

(SEMAAN, 2013). A próxima subseção irá descrever este algoritmo.

2.7.4.3.1 DBSCAN

O algoritmo DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) foi criado com o intuito de descobrir diferentes formas de agrupamentos (Figura 7) e indicar a presença de ruído nas bases de dados (ESTER et al., 1996). Ele utiliza o conceito de densidade baseada em centro, ou seja, a densidade de um ponto no conjunto de dados é a quantidade de pontos dentro de um raio de vizinhança (ε), incluindo o próprio ponto. Este algoritmo contém dois parâmetros de entrada, o raio (ε) e a quantidade mínima de pontos em um determinado raio (*minPts*) (SEMAAN, 2013).

O fato de densidade baseada no centro, possibilita classificar um ponto como estando: no interior de uma região densa (ponto de centro), no limite de uma região densa (ponto de limite) ou em uma região ocupada esporadicamente (ponto de ruídos), conforme ilustrado na Figura 8a. Dado essas classificações, o DBSCAN pode ser descrito pelo algoritmo 3 (TAN; STEINBACH; KUMAR, 2009).

Para alcançar resultados de boa qualidade, é necessário ajustar o parâmetro do raio (ε) para que seu valor não seja tão alto de modo que, todos os pontos tenham densidade n (solução com apenas um grupo), e nem tão baixo em que todos os objetos tenham

densidade 1 (solução com n grupos denominados *singletons*) (SEMAAN, 2013).

Algoritmo 3: Algoritmo DBSCAN

início

- Rotular todos os pontos como de centro, limite ou ruído;
- Eliminar os pontos de ruído;
- Colocar uma aresta entre todos os pontos de centro que estejam dentro do raio ε uns dos outros;
- Tornar cada grupo de pontos de centro conectados em grupo separado;
- Atribuir cada ponto de limite a um dos grupos dos seus pontos de centro associados;

fim

Fonte – (TAN; STEINBACH; KUMAR, 2009)

Na Figura 8b é mostrado o resultado obtido pela execução do DBSCAN, no qual pode-se observar que, pontos de centro e de limite formam grupos, enquanto o de ruídos permanecem afastados.

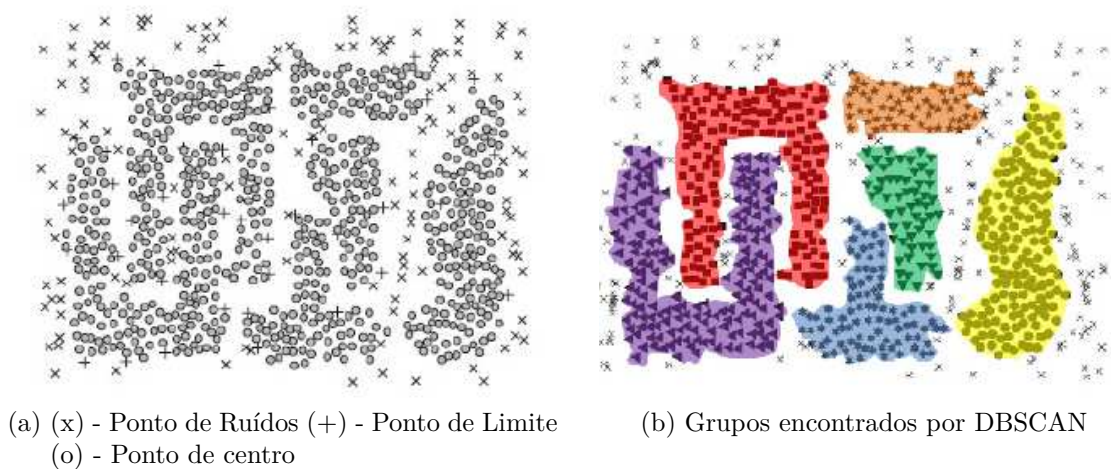


Figura 8 – Agrupamento DBSCAN para 3.0000 pontos bidimensionais

Fonte – adaptado de (TAN; STEINBACH; KUMAR, 2009)

Pelo fato do algoritmo DBSCAN ser baseado em densidade, ele pode encontrar muitos grupos que o *k-means* não poderia encontrar. Portanto, ele também possui suas desvantagens: tem dificuldade de trabalhar com grupos de densidades muito variadas e pode ser custoso calcular os pontos vizinhos quando requer o cálculo de proximidade entre pares (TAN; STEINBACH; KUMAR, 2009).

2.8 Validação de Agrupamento

A validação de agrupamento aborda os processos formais que avaliam, de forma objetiva e quantitativa, os resultados da análise do agrupamento. Os índices de validade de agrupamento, ou medidas de avaliação, podem ser estabelecidos a partir de três diferentes tipos: externo, relativo e interno (JAIN; DUBES, 1988)

Índices baseados em critérios externos usam informações que não estão contidas no conjunto de dados (TAN; STEINBACH; KUMAR, 2009). Um exemplo muito conhecido de índice externo é o *Rand Index* (RI), e o mesmo é dado pela equação 2.3. As variáveis U e V representam duas matrizes de agrupamento particional exclusivo, onde uma seria um agrupamento gerado e a outra a solução ideal. As demais variáveis são definidas como (HORTA, 2013):

- a : número de pares de dados pertencentes aos mesmos grupos tanto em U quanto em V ;
- b : número de pares de dados pertencentes a grupos diferentes em U mas nos mesmos grupos em V ;
- c : número de pares de dados pertencentes aos mesmos grupos em U mas em diferentes grupos em V ;
- d : número de pares de dados pertencentes a grupos diferentes tanto em U quanto em V .

$$RI(U, V) = \frac{a + d}{a + b + c + d} \quad (2.3)$$

Já o propósito dos índices relativos, é comparar diferentes agrupamentos ou grupos, sejam supervisionados ou não. Como exemplo, dois agrupamento obtidos pelo *k-means* podem ser comparados utilizando o método de entropia (TAN; STEINBACH; KUMAR, 2009). Este método analisa o quão distantes as classes de dados estão dentro de um grupo, calculando primeiramente a sua distribuição, ou seja, para uma classe j é calculado seu p_{ij} , que representa a probabilidade de um objeto do grupo i pertencer à classe j . O p_{ij} é calculado como $p_{ij} = m_{ij}/m_i$, o qual m_{ij} é a quantidade de objetos da classe j no grupo i e, m_i é a quantidade de objetos no grupo i (GROSS, 2014).

Então, a entropia de cada grupo i é calculada pela equação 2.4, onde L é o número de classes (GROSS, 2014). Quanto maior a entropia de um registro, mais semelhante é a partilha dos seus valores (CASTANHEIRA, 2008).

$$e_i = - \sum_{j=1}^L p_{ij} \log_2 p_{ij} \quad (2.4)$$

Para finalizar, os índices internos qualificam o agrupamento utilizando somente informações do conjunto de dados. Dentre esses índices, merece destaque o índice de Silhueta (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010), representado pela equação 2.5.

$$Silhueta(x_i) = \frac{b(x_i) - a(x_i)}{\max[a(x_i), b(x_i)]} \quad (2.5)$$

Considere um dado x_i pertencente a um grupo C_a . A distância média de x_i para os demais dados de C_a é representada por $a(x_i)$. Levando em consideração outro grupo C_c , a distância média do dado x_i para todos os dados do grupo C_c será referenciada por $d(x_i, C_c)$. Após realizar o cálculo de $d(x_i, C_c)$ para todos os grupos $C_c \neq C_a$, é selecionado o menor valor, representado pela equação 2.6. Este valor ($b(x_i)$) representa a distância de x_i para o grupo mais próximo. A silhueta depende do cálculo de todas as distâncias entre os dados da base, exigindo complexidade $O(N^2)$, o que o torna custosa computacionalmente (ALVES et al., 2007).

$$b(x_i) = \min[d(x_i, C_c)], C_c \neq C_a \quad (2.6)$$

Para solucionar essa limitação, a silhueta simplificada (VENDRAMIN; CAMPELLO; HRUSCHKA, 2010) é uma opção interessante. Ela determina a qualidade de um agrupamento calculando a similaridade (obtidas pelas medidas de distâncias) entre os dados de um grupo e a distância desses dados ao centroide do grupo mais próximo. Dessa forma, pode-se identificar quais dados estão bem posicionados em seus devidos grupos e quais devem estar em outro grupo (CONCEIÇÃO et al., 2015). O valor de $a(x_i)$ da equação 2.5 torna-se a distância do dado x_i ao centroide do seu grupo (C_a) e no lugar do cálculo de $d(x_i, C_c)$ como a distância média do dado x_i para todos os dados de C_c , $C_c \neq C_a$, somente a distância entre x_i e o centroide de C_c é calculada, reduzindo a complexidade para $O(N)$ (ALVES et al., 2007).

2.9 Trabalhos Relacionados

Nesta seção serão apresentados os trabalhos relacionados ao tema desta pesquisa. A quantidade de trabalhos encontrados foi muito pequena, isto se deve ao fato de que, os dados e estatísticas do MPT se encontravam perdidos em banco de dados governamentais ou em anuários pouco compreensíveis, o que dificultava a pesquisa sobre o assunto e sua compreensão (OBSERVATÓRIO, 2017).

Em Bartolomeu et al. (2002), os autores aplicaram técnicas de mineração de dados para guiar as tomadas de decisões das políticas de controle e prevenção de acidentes e doenças ocupacionais. Desse modo, eles descobriram informações significativas sobre os

acidentes comunicados ao INSS, porém, o trabalho foi voltado somente para o estado de Santa Catarina (SC).

Já um estudo publicado por [Pignata \(2016\)](#), foi utilizado técnicas de agrupamento particionais nos dados do Anuário Estatístico de Acidentes de Trabalho, o qual possuem informações sobre CATs por localizações geográficas dos acontecimentos. O autor chegou a conclusão que nas regiões que são mais industrializadas e que possuem maior atividade econômica, são onde mais ocorrem acidentes de trabalho. No entanto, ele utilizou somente o algoritmo *k-means* para chegar a esta conclusão, não foi aplicado outros tipos de agrupamento para ver se os resultados seriam semelhantes.

[Guimarães et al. \(2000\)](#), também optaram por ferramentas de MD como estratégia para instituições públicas, mas neste caso, foi voltado para o sistema transacional do Ministério Público de Rondônia, chamado Controle de Inquéritos Policiais (CIPO). Algoritmo de árvore de decisão e ferramentas baseadas em indução de regras, foram utilizadas para constatar que existe conhecimento aproveitável em base de dados de instituições públicas, fazendo com que o Ministério Público de Rondônia aderisse critérios que, produzirão o ambiente necessário para utilização de ferramentas de MD, integradas aos sistemas tradicionais de tomada de decisão.

Por último, foi criado no dia 28 de abril de 2017, o observatório digital de saúde e segurança do trabalho. As visões que mais se destacam nesta ferramenta são os indicadores de incidência, localização geográfica, gastos previdenciários acumulados, mortes acidentárias, número de notificações de acidentes (CATs) e diversos tipos de gráficos ([OBSERVATÓRIO, 2017](#)). No entanto, o observatório contribui apenas para visualização e sumarização dos dados dos acidentes de trabalho, não utilizando técnicas de mineração de dados e nem inteligência artificial.

2.10 Considerações Finais

Neste capítulo foi apresentado uma visão geral sobre acidentes de trabalho, CAT, benefícios previdenciários e como o MPT gerencia os acidentes, além de apresentar todo o processo de KDD, em especial as etapas de pré-processamento e mineração de dados, as quais serão utilizadas neste trabalho.

O capítulo 3 apresentará com maior detalhe como este trabalho foi desenvolvido, explicando cada técnica de pré-processamento, agrupamento, validação de agrupamento e ferramentas utilizadas, assim como as bases de dados geradas.

3 Desenvolvimento

3.1 Introdução

Neste capítulo será apresentado como o trabalho foi desenvolvido. A seção 3.2 detalha a base de dados utilizada, apresentando todos seus atributos e os diferentes valores que cada atributo pode assumir. A seção 3.3 descreve todo o procedimento de pré-processamento realizado. A seção 3.4 apresenta as bases de dados geradas a partir do pré-processamento e que serão utilizadas neste trabalho. A seção 3.5 mostra a ferramenta, as configurações dos algoritmos de agrupamento utilizados e a medida de validação escolhida para avaliar os agrupamentos. Por último, a seção 3.6 traz as considerações finais.

3.2 Apresentação da base de dados do MPT

O Observatório Digital de Saúde e Segurança do Trabalho disponibilizou dois conjuntos de dados, um de acidentes de trabalho notificados e outro de benefícios previdenciários concedidos aos trabalhadores. O conjunto utilizado nesta trabalho foi o de benefícios, sendo que, possui um grupo de alunos trabalhando nessas bases e a base de acidentes está sendo desenvolvido por um outro aluno do grupo.

O conjunto de benefícios previdenciários contém dados entre os anos de 2012 a 2017 e foram retirados do Sistema Único de Informações de Benefícios da Previdência Social (SISBEN). Possui 1.327.958 instâncias e 14 atributos, os quais 5 são do tipo numérico (ano do benefício, idade do beneficiário, espécie, despesa total e dias perdidos) e os demais categóricos. A Tabela 1 detalha cada atributo trazendo o nome, a sua descrição, o tipo e a quantidade de valores diferentes (Qtd.), bem como a quantidade de valores ausentes e não aplicáveis (NA), existentes no conjunto de dados.

Como pode ser visto na Tabela 1, este conjunto de dados possui muitos valores ausentes e não aplicáveis, além de conter muitos atributos categóricos. Para a tarefa de agrupamento, a maioria dos algoritmos trabalham apenas com valores numéricos e não conseguem lidar com valores ausentes. A seção 3.3 detalha os pré-processamentos realizados afim de transformar a base para um formato que possa ser usado pelos algoritmos de agrupamento.

Tabela 1 – Descrição dos dados do conjunto de benefícios previdenciários

Atributo	Descrição	Tipo	Qtd.	Ausente	NA
Código da categoria CID	A Classificação Internacional de Doenças (CID) trata-se de um catálogo que fornece uma codificação padrão relativos à classificação de doenças e de uma grande variedade de sintomas (ICLINIC, 2016). Este atributo representa o código que se enquadra o problema de uma determinada instância.	CHAR(3)	1.364	0	12.395
Nome da categoria CID	Representa o nome da categoria CID de uma determinada instância.	STRING	1.364	0	12.396
Descrição da classe do CNAE	A Classificação Nacional de Atividades Econômicas (CNAE) é um meio de padronizar os códigos de atividades econômicas e os critérios de enquadramento utilizados pelos diferentes órgãos da administração tributária do Brasil. Ela é estruturada em seção, divisão, grupo e classe (KRUMHEUER, 2017). Este atributo representa a descrição da classe CNAE da empresa de uma determinada instância.	STRING	668	141.305	0
Município do Benefício	Representa o município da residência, empresa, local do acidente ou requerimento de uma determinada instância.	STRING	5.285	0	8.094
UF do Benefício	Representa a UF de uma determinada instância.	STRING	28	0	8.094
Ano do Benefício	Representa o ano em que foi concedido o benefício de uma determinada instância.	SMALLINT	5	0	0
Idade do Beneficiário	Representa a idade da pessoa que foi beneficiada.	SMALLINT	104	0	0
Sexo do Trabalhador	Representa o sexo da pessoa que foi beneficiada.	STRING	2	0	0

continua

Espécie	Representa a espécie do benefício concedido. Neste conjunto de dados só contém a espécie 91, que é o Auxílio-doença acidentário.	SMALLINT	1	0	0
Despesa Total	Representa o valor total gasto com o benefício de uma determinada instância.	NUMERIC(15,10)	1.163.071	0	0
Dias perdidos	Representa a quantidade de dias de trabalho perdido referente ao trabalhador de uma determinada instância.	INTEGER	3.517	0	0
Agrupamento da categoria CID	Este atributo foi criado pelo MPT para agrupar as categorias CID que são similares. Ele representa qual o agrupamento da categoria CID se encontra uma determinada instância.	STRING	37	12.395	0
Agrupamento da categoria CID - Simplificado	Este atributo é a parte mais genérica do atributo anterior. O benefício concedido em uma determinada instância, pode ser originado devido ao trabalhador sofrer um acidente, doença ou outras causas, além de possuir algumas instâncias com esse atributo ausente, que são os quatro possíveis valores desse atributo.	STRING	4	12.395	0
Agrupamento da categoria CID - Doença	Este atributo representa um agrupamento da categoria CID para doenças. Caso uma determinada instância não conter valor nesse atributo, indica que a mesma foi originada por um acidente ou por outras causas.	STRING	114	856.255	0

conclusão

3.3 Pré-processamentos realizados

Para a tarefa de pré-processamento, o especialista de domínio foi consultado afim de tirar dúvidas, indicar caminhos e complementar informações referente a base de dados. Feito isso, foi decidido remover do conjunto de dados os atributos desnecessários à tarefa de agrupamento ou aqueles que representam informação duplicada. Os atributos código da categoria CID e nome da categoria CID foram removidos por serem representados pelo atributo agrupamento da categoria CID, o atributo UF do benefício por ser representado pelo Município do benefício e a espécie por conter somente um valor, conforme descrito na Tabela 1.

Como a maioria das técnicas de agrupamento trabalham com dados numéricos, a segunda etapa foi realizar a conversão dos dados categóricos. Para isso utilizou-se a codificação inteira-binária e a codificação 1-de-n.

Na codificação inteira-binária, deve-se associar a cada valor de um atributo, um inteiro no intervalo de $[0, m-1]$, mantendo a ordem no caso do valor ser ordinal (que estabelecem uma ordem ou hierarquia). Feito isso, é realizado a conversão de cada um dos m inteiros para binário. São necessários $n = \lceil \log_2(m) \rceil$ dígitos binários para representar estes números inteiros, os quais serão retratados por n atributos. A Tabela 2 ilustra um exemplo de codificação inteira-binária. Um problema desta codificação é a criação de relacionamentos não pretendidos entre os atributos convertidos, por exemplo, os atributos x_2 e x_3 da Tabela 2, são correlacionados para o valor “Bom” (TAN; STEINBACH; KUMAR, 2009).

Tabela 2 – Codificação inteira-binária de um atributo categorizado

Valor Categorizado	Valor Inteiro	x_1	x_2	x_3
Terrível	0	0	0	0
Fraco	1	0	0	1
Satisfatório	2	0	1	0
Bom	3	0	1	1
Excelente	4	1	0	0

Fonte – (TAN; STEINBACH; KUMAR, 2009)

Já na codificação 1-de-n é criado um atributo para cada valor categórico. O atributo que corresponde ao valor deve ser preenchido com 1 e os demais com 0. A Tabela 3 mostra um exemplo de conversão utilizando codificação 1-de-n. O problema desta codificação é que se em um atributo de uma base possuir muitos valores diferentes, será necessário criar muito atributos, o que aumentará a base de dados e poderá perder a qualidade e eficiência do agrupamento.

Tabela 3 – Codificação 1-de-n de um atributo categorizado

Valor Categorizado	Valor Inteiro	x_1	x_2	x_3	x_4	x_5
Terrível	0	1	0	0	0	0
Fraco	1	0	1	0	0	0
Satisfatório	2	0	0	1	0	0
Bom	3	0	0	0	1	0
Excelente	4	0	0	0	0	1

Fonte – (TAN; STEINBACH; KUMAR, 2009)

Como cada codificação tem suas vantagens e desvantagens, foram criadas diferentes bases aplicando os dois tipos de codificação, as quais serão apresentadas na seção 3.4. Porém, antes de aplicá-las, foi necessário converter dois atributos para uma versão mais resumida, pois continham muitos valores diferentes, o que implicaria na criação de um número muito grande de atributos na base pré-processada.

O primeiro atributo é a descrição da classe do CNAE. Conforme descrito na Tabela 1, o CNAE é estabelecido em uma hierarquia. Na base de dados de benefícios, este atributo era representado pela classe, o qual possuía 668 valores diferentes, contando com o ausente. O atributo CNAE classe foi convertido para CNAE seção, que possui apenas 21 valores diferentes representados pelas letras *A* até *U* respectivamente. Então, para converter em um número inteiro e aplicar as codificações, foram enumerados de 1 até 22, onde o 22 representa o valor ausente.

Na Figura 9 é exemplificado uma conversão do valor CNAE. Nesta figura, o valor do atributo Descrição da classe do CNAE na primeira instância do conjunto de dados é Administração pública em geral, representado pelo código 8411-6. Este atributo está dentro do grupo “841 - Administração do estado e da política econômica e social” que faz parte da divisão “84 - Administração pública, defesa e seguridade social” e que pertence à seção “O - Administração pública, defesa e seguridade social”. Como a letra “O” é a décima quinta letra do alfabeto, então este atributo será convertido para o valor 15.

Hierarquia		
Seção:	O	ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL
Divisão:	84	ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL
Grupo:	841	ADMINISTRAÇÃO DO ESTADO E DA POLÍTICA ECONÔMICA E SOCIAL
Este grupo contém as seguintes classes:		
8411-6	ADMINISTRAÇÃO PÚBLICA EM GERAL	
8412-4	REGULAÇÃO DAS ATIVIDADES DE SAÚDE, EDUCAÇÃO, SERVIÇOS CULTURAIS E OUTROS SERVIÇOS SOCIAIS	
8413-2	REGULAÇÃO DAS ATIVIDADES ECONÔMICAS	

Figura 9 – Exemplo de hierarquia do CNAE.

Fonte – (IBGE, 2018)

O segundo atributo convertido para uma representação com menos valores possí-

veis foi o Município do benefício, que possui 5.285 valores diferentes na base de dados. Ele foi convertido pelos códigos das Mesorregiões, que são subdivisões dos estados brasileiros que agrupam diversos municípios de uma área geográfica com semelhanças econômicas e sociais (EDUCAÇÃO, 2010). O Brasil possui 137 Mesorregiões. Além disso, o DATASUS (Departamento de Informática do SUS) (DATASUS, 2018) disponibiliza uma tabela que contém informações referentes a outras 27 Mesorregiões pertencentes aos Municípios ignorados (um para cada Unidade de Federação e outro para o exterior) e uma para os Municípios transferidos de Goiás para Tocantins. Assim, este trabalho considerou a existência de 165 Mesorregiões, as quais foram convertidas para 165 atributos.

Na Figura 10 é ilustrado um exemplo da tabela de Mesorregiões do DATASUS. Pode-se observar pelo registro selecionado, que o município Barbacena pertence à mesorregião Campo das Vertentes. O valor do atributo Município do benefício na primeira instância do conjunto de dados é Barbacena, portanto, o mesmo será convertido para o código da mesorregião Campo das Vertentes, que é 3111.

A	B	C	D	E	F	G
Ano	Código da Unidade da Federação	Nome da Unidade da Federação	Código do Município	Nome do Município	Região Metropolitana	Código da Mesorregião
2015	31	Minas Gerais	3105608	Barbacena		3111
2015	31	Minas Gerais	3105707	Barra Longa		3112
2015	31	Minas Gerais	3105905	Barroso		3111
2015	31	Minas Gerais	3106002	Bela Vista de Minas		3107
2015	31	Minas Gerais	3106101	Belmiro Braga		3112
2015	31	Minas Gerais	3106200	Belo Horizonte	RM Belo Horizonte	3107
2015	31	Minas Gerais	3106309	Belo Oriente	RM Vale do Aço	3108

Figura 10 – Exemplo da planilha de Mesorregiões do DATASUS.

Fonte – (DATASUS, 2018)

O atributo sexo, que possuía somente os valores feminino e masculino, foi convertido para 1 e 0 respectivamente.

Como os atributos ano do benefício, idade do beneficiário, despesa total e dias perdidos estão numa escala bem diferentes dos demais, eles poderiam afetar o cálculo de medidas de distância, que são comumente usadas em algoritmos de agrupamento. Assim, a fim de evitar que alguns atributos tenham mais impacto que outros, todos eles foram reescalados para o intervalo $[0,1]$. A equação utilizada para reescalar estes atributos foi a 3.1, onde d indica o atributo a ser reescalado, min_d o menor valor da coluna do atributo e max_d o maior valor.

$$d' = \frac{(d - \min_d)}{(\max_d - \min_d)} \quad (3.1)$$

A última etapa realizada foi a de converter a base para o formato ARFF (Formato de Arquivo de Relação de Atributos). Um arquivo ARFF é um arquivo de texto ASCII que descreve uma lista de instâncias que compartilham um conjunto de atributos, o qual foi desenvolvido para uso com o *software* de aprendizado de máquina Weka (WAIKATO, 2008), que será detalhado na subseção 3.5.

Um exemplo de arquivo no formato ARFF pode ser observado na Figura 11. Ele é dividido em duas seções: cabeçalho e dados. O cabeçalho contém o nome da relação (linha 3), uma lista dos atributos e seus tipos (linha 5 à 22). Os dados são separados por vírgulas e começam após a declaração @data (linha 25 à 34). Linhas que começam com % são comentários (linha 1).

```

1 % Base inteira-binária
2
3 @relation Dados_beneficio
4
5 @attribute ds_cnae_classe_beneficio_1 numeric
6 @attribute ds_cnae_classe_beneficio_2 numeric
7 @attribute ds_cnae_classe_beneficio_3 numeric
8 @attribute ds_cnae_classe_beneficio_4 numeric
9 @attribute ds_cnae_classe_beneficio_5 numeric
10 @attribute nm_municipio_1 numeric
11 @attribute nm_municipio_2 numeric
12 @attribute nm_municipio_3 numeric
13 @attribute nm_municipio_4 numeric
14 @attribute nm_municipio_5 numeric
15 @attribute nm_municipio_6 numeric
16 @attribute nm_municipio_7 numeric
17 @attribute nm_municipio_8 numeric
18 @attribute ano_beneficio numeric
19 @attribute idade_beneficio numeric
20 @attribute cd_tipo_sexo_empregado_beneficio numeric
21 @attribute qt_despesa_total numeric
22 @attribute qt_dias_perdidos numeric
23
24 @data
25 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.5,0.53125,0,0.028922,0.043461
26 0,0,0,0,1,0,0,0,0,0,0,0,0,1,0.75,0.458333,0,0.042041,0.072035
27 0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0.25,0.75,0,0.048001,0.102449
28 0,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0.5,0.239583,1,0.074154,0.108532
29 0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0.75,0.59375,0,0.022724,0.034417
30 0,0,0,1,1,0,0,0,0,0,0,1,0,1,0,0.75,0.729167,0,0,0
31 0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,0.25,0.65625,1,0.003416,0.007364
32 0,0,0,1,0,0,0,0,0,0,0,0,1,0,1,0.5,0.5625,1,0.093515,0.196814
33 0,0,1,0,1,0,0,0,0,0,0,0,1,0,0.75,0.53125,1,0.03898,0.031055
34 0,0,1,1,0,0,0,0,0,0,0,1,1,1,1,0.802083,0,0.00185,0.001761

```

ferramenta criada para este propósito. Na Figura 12 é ilustrado a tela inicial da ferramenta. As funções codificação 1-de-n, codificação inteira-binária, separar base de dados, remover coluna e reescalar, foram criadas de uma forma genérica para serem aplicadas a qualquer conjunto de dados, desde que estejam representados em uma planilha com a extensão *.csv*. Já as demais funções, foram criadas especificamente para o conjunto de dados estudado neste trabalho. Esta ferramenta foi feita na linguagem Java e seu uso é bem simples. Para as funcionalidades que são genéricas, basta indicar o caminho da base de dados e a coluna que queira converter ou remover, enquanto que para as específicas, deve ser indicado também o caminho das planilhas que serão disponibilizadas juntamente com a ferramenta e que permite fazer a conversão dos atributos.

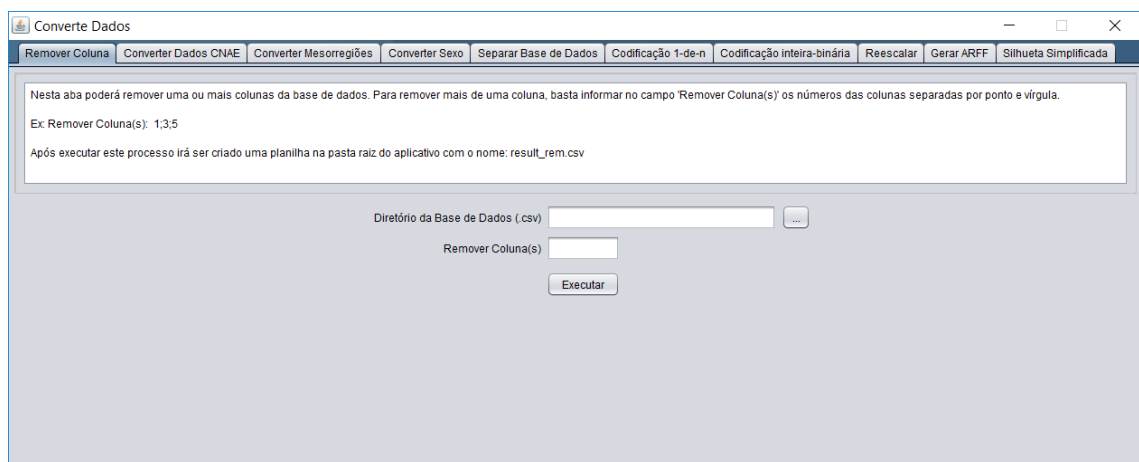


Figura 12 – Tela da ferramenta desenvolvida neste trabalho para pré-processamento dos dados.

3.4 Bases de dados geradas

Depois de aplicar o pré-processamento, foi obtida uma base de dados contendo as mesmas 1.327.958 instâncias, sendo 319 atributos considerando a conversão 1-de-n e 34 atributos considerando a conversão inteira-binária. No entanto, trabalhar com uma base com tantas instâncias e atributos, pode não ser adequado quando se tem um trabalho exploratório que está em busca de encontrar grupos interessantes para o problema.

Devido a isto, foi selecionado o atributo Agrupamento da categoria CID - Doença (conforme descrito na Tabela 1, quando preenchido indica que a determinada instância foi originada de uma doença, caso contrário de um acidente) e usado para separar a base em duas, gerando assim uma base referente a doenças com 471.703 instâncias e outra referente a acidentes com 856.255 instâncias. Como a base de acidentes possui o mesmo valor no atributo Agrupamento da categoria CID - Doença, que é o ausente, o mesmo foi removido.

A fim de explorar as técnicas de agrupamento em diferentes visões e visto que as

bases geradas ainda estavam com muitas instâncias, cada uma delas foi redividida pelo atributo ano do benefício, gerando assim as seguintes bases de dados:

- Base_2012_1n_D: referente a doença, ano 2012, aplicado a codificação 1-de-n, 111.357 instâncias e 296 atributos.
- Base_2013_1n_D: referente a doença, ano 2013, aplicado a codificação 1-de-n, 112.804 instâncias e 296 atributos.
- Base_2014_1n_D: referente a doença, ano 2014, aplicado a codificação 1-de-n, 101.059 instâncias e 296 atributos.
- Base_2015_1n_D: referente a doença, ano 2015, aplicado a codificação 1-de-n, 69.120 instâncias e 296 atributos.
- Base_2016_1n_D: referente a doença, ano 2016, aplicado a codificação 1-de-n, 77.363 instâncias e 296 atributos.
- Base_2012_1n_A: referente a acidente, ano 2012, aplicado a codificação 1-de-n, 196.233 instâncias e 186 atributos.
- Base_2013_1n_A: referente a acidente, ano 2013, aplicado a codificação 1-de-n, 194.623 instâncias e 186 atributos.
- Base_2014_1n_A: referente a acidente, ano 2014, aplicado a codificação 1-de-n, 182.107 instâncias e 186 atributos.
- Base_2015_1n_A: referente a acidente, ano 2015, aplicado a codificação 1-de-n, 129.759 instâncias e 186 atributos.
- Base_2016_1n_A: referente a acidente, ano 2016, aplicado a codificação 1-de-n, 153.533 instâncias e 186 atributos.
- Base_2012_IntBin_D: referente a doença, ano 2012, aplicado a codificação inteira-binária, 111.357 instâncias e 31 atributos.
- Base_2013_IntBin_D: referente a doença, ano 2013, aplicado a codificação inteira-binária, 112.804 instâncias e 31 atributos.
- Base_2014_IntBin_D: referente a doença, ano 2014, aplicado a codificação inteira-binária, 101.059 instâncias e 31 atributos.
- Base_2015_IntBin_D: referente a doença, ano 2015, aplicado a codificação inteira-binária, 69.120 instâncias e 31 atributos.
- Base_2016_IntBin_D: referente a doença, ano 2016, aplicado a codificação inteira-binária, 77.363 instâncias e 31 atributos.

- Base_2012_IntBin_A: referente a acidente, ano 2012, aplicado a codificação inteira-binária, 196.233 instâncias e 24 atributos.
- Base_2013_IntBin_A: referente a acidente, ano 2013, aplicado a codificação inteira-binária, 194.623 instâncias e 24 atributos.
- Base_2014_IntBin_A: referente a acidente, ano 2014, aplicado a codificação inteira-binária, 182.107 instâncias e 24 atributos.
- Base_2015_IntBin_A: referente a acidente, ano 2015, aplicado a codificação inteira-binária, 129.759 instâncias e 24 atributos.
- Base_2016_IntBin_A: referente a acidente, ano 2016, aplicado a codificação inteira-binária, 153.533 instâncias e 24 atributos.

3.5 Métodos de agrupamento e medida de validação utilizados

Para executar os métodos de agrupamento, foi utilizado o *software* Weka (EIBE; MARK; IAN, 2016). Esse *software* foi desenvolvido na linguagem Java, pelo curso de Ciência da Computação da Universidade de Waikato na Nova Zelândia. O Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados, o qual contém ferramentas para pré-processamento, agrupamento, classificação, associação, regressão e visualização (WITTEN; FRANK; HALL, 2011).

Os algoritmos de agrupamento utilizados neste trabalho foram o *k-means*, *Canopy* e EM. Ambos foram executados alterando o número de grupos de 2 à 10, deixando as demais configurações com os valores padrões. Esses algoritmos de agrupamento foram escolhidos por serem simples, eficientes e bastante utilizados. Experimentos usando o algoritmo DBSCAN estão em andamento. Como os resultados estão demorando muito para serem processados, eles não serão descritos neste trabalho.

Os resultados obtidos pelos algoritmos utilizados poderão fornecer direcionamentos futuros do que pode ser explorados nestes dados e de quais outros algoritmos merecem ser investigados.

A medida de validação usada neste trabalho foi a Silhueta Simplificada, a qual foi detalhada na seção 2.8. O principal motivo dessa escolha é que esta validação possui complexidade computacional linear $O(n)$ em relação ao número de elementos no conjunto de dados, o que faz com que tenha um melhor desempenho computacional (CONCEIÇÃO et al., 2015), além de ter sido utilizada em diferentes trabalhos de agrupamento como (NASSIF, 2011; BORGES, 2010; FERREIRA, 2012; CONCEIÇÃO et al., 2015), entre outros.

3.6 Considerações Finais

Este capítulo descreveu em detalhes a base de dados do MPT que foi estudada e todos os pré-processamentos necessários para realizar a limpeza da mesma. Também foi informado quais os agrupamentos utilizados e suas configurações, além do método de validação de agrupamento.

O capítulo 4 apresentará os resultados deste trabalho utilizando as propostas que foram descritas neste capítulo e também irá discutir os resultados obtidos a partir dos experimentos.

4 Resultados

4.1 Introdução

Este capítulo tem como objetivo analisar os resultados obtidos usando os métodos de agrupamento e validação de agrupamento descritos no Capítulo 3. Os resultados dos agrupamentos são avaliados utilizando o índice de Silhueta Simplificada, de modo que quanto mais próximo o valor estiver de 1 melhor é o resultado. Esses resultados serão apresentados em tabelas e, com o objetivo de uma melhor visualização dos mesmos, para cada tabela existe um gráfico de barras associado.

A seção 4.2 apresenta os resultados do algoritmo *k-means*, a seção 4.3 do algoritmo *canopy* e a seção 4.4 do algoritmo EM. A seção 4.5 traz as considerações finais.

4.2 Algoritmo *k-means*

As Tabelas 4 e 5 apresentam os valores da silhueta simplificada das bases de doenças e as Tabelas 6 e 7 das bases de acidentes, ambas agrupadas pelo algoritmo *k-means*. As linhas das tabelas indicam os diferentes valores de k utilizados e as colunas indicam os subconjuntos de dados extraídos da base de acordo com o ano do benefício.

As bases das Tabelas 4 e 6 foram pré-processadas usando codificação inteira-binária enquanto as bases das Tabelas 5 e 7 pela codificação 1-de-n. Tanto nas bases de doenças quanto na de acidentes percebe-se que a medida que vai aumentando o número de grupos há pouca modificação na qualidade do agrupamento. Pode-se concluir que o número de grupos tem pouca influência na qualidade do resultado do agrupamento. Nas Figuras 13 e 14 são ilustrados graficamente o valor da silhueta simplificada usando o algoritmo *k-means* com diferentes valores de k nas bases de doenças e acidentes respectivamente.

Tabela 4 – Agrupamento *k-means*, Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.

N° de agrupamento	2012	2013	2014	2015	2016
2	0,12602744	0,117261836	0,127150733	0,117988416	0,120514
3	0,150733864	0,12308381	0,149893253	0,12225193	0,121256312
4	0,112569302	0,112415943	0,134077007	0,135581155	0,160328634
5	0,133674537	0,133370751	0,130017952	0,144040642	0,137207953
6	0,145877747	0,147086842	0,124730867	0,139956092	0,134026321
7	0,14966364	0,159473118	0,153079033	0,130173987	0,145810393
8	0,143486225	0,154643519	0,147522572	0,138332955	0,162486255
9	0,154505373	0,15754464	0,153834635	0,147725192	0,15080893
10	0,150314436	0,152083899	0,160975968	0,16620386	0,156103109

Tabela 5 – Agrupamento *k-means*, Base_ano_1n_D, onde o ano vai de 2012 à 2016.

N° de agrupamento	2012	2013	2014	2015	2016
2	0,175938026	0,174830249	0,175148706	0,135384398	0,177203484
3	0,150812995	0,177902426	0,139793084	0,148923925	0,171000241
4	0,147381957	0,193853923	0,155419924	0,149995046	0,14755281
5	0,168024144	0,141984178	0,144743137	0,160872325	0,15980839
6	0,169747846	0,169716403	0,165332023	0,15439788	0,169243309
7	0,164441878	0,165749814	0,16638014	0,179172294	0,152715145
8	0,167978652	0,172503636	0,143235678	0,162345548	0,155316876
9	0,162020018	0,157569407	0,17135167	0,170919762	0,156551076
10	0,159518901	0,163208669	0,163439832	0,173067683	0,156253443

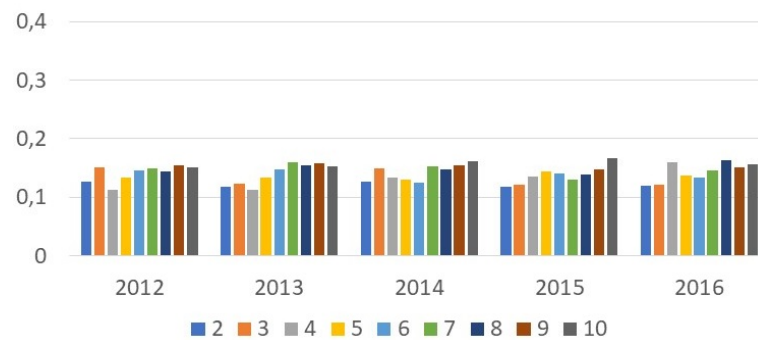
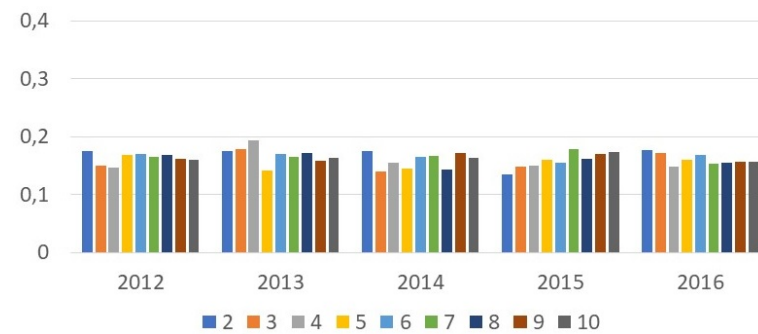
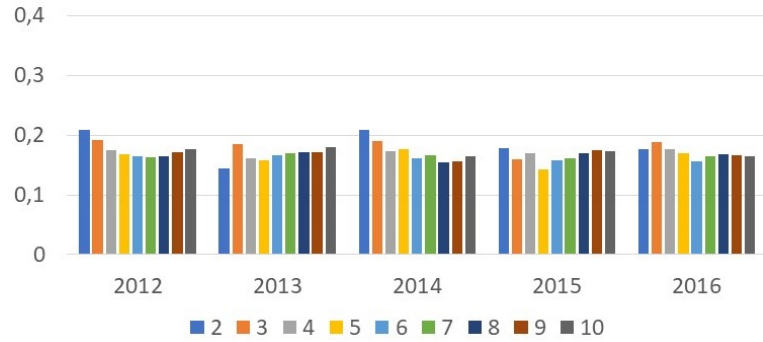
(a) Agrupamento *k-means*, Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.(b) Agrupamento *k-means*, Base_ano_1n_D, onde o ano vai de 2012 à 2016.Figura 13 – Gráficos das bases de dados de doenças utilizando o algoritmo *k-means* e modificando o número de agrupamento.

Tabela 6 – Agrupamento *k-means* - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.

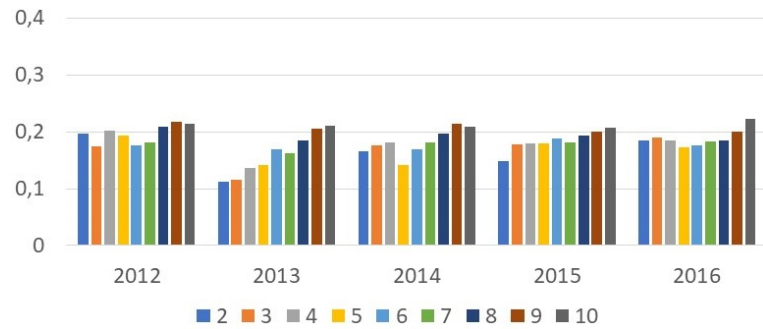
N° de agrupamento	2012	2013	2014	2015	2016
2	0,208878945	0,145068742	0,208541447	0,178636551	0,176299301
3	0,191868277	0,185303138	0,190098467	0,159662917	0,18806096
4	0,174038193	0,160510718	0,17376876	0,169569563	0,177091219
5	0,167492426	0,158137356	0,17626724	0,143296137	0,169344266
6	0,163953219	0,165701427	0,160678617	0,15741293	0,156802413
7	0,162414413	0,169551325	0,166341905	0,161804316	0,165287821
8	0,1643975	0,171562068	0,155201681	0,169129245	0,16753939
9	0,171040113	0,171550668	0,155359315	0,175146352	0,165752519
10	0,176018544	0,180475292	0,163977181	0,173575979	0,163945357

Tabela 7 – Agrupamento *k-means* - Base_ano_1n_A, onde o ano vai de 2012 à 2016.

N° de agrupamento	2012	2013	2014	2015	2016
2	0,196344355	0,112767597	0,166582784	0,148394132	0,184320658
3	0,174692926	0,116294809	0,176621408	0,178155275	0,189150923
4	0,201193558	0,136751702	0,180860654	0,179001882	0,185244382
5	0,192849755	0,141377364	0,142399633	0,179139337	0,172132024
6	0,176538071	0,169201109	0,169275607	0,187680156	0,175274851
7	0,180545387	0,162227011	0,181792395	0,181317171	0,182883752
8	0,208618747	0,184801973	0,197084634	0,194064509	0,183983263
9	0,216797123	0,206040355	0,213515077	0,200242345	0,200217254
10	0,214740256	0,210269269	0,208301228	0,206430132	0,222466389



(a) Agrupamento *k-means* - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.



(b) Agrupamento *k-means* - Base_ano_1n_A, onde o ano vai de 2012 à 2016.

Figura 14 – Gráficos das bases de dados de acidentes utilizando o algoritmo *k-means* e modificando o número de agrupamento.

Pode-se observar que o agrupamento gerado pelo *k-means* usando a base com codificação 1-de-n é ligeiramente melhor que os resultados usando a codificação inteira-binária, tanto na base de acidentes quanto na base de doenças.

No entanto, os maiores valores de silhueta obtidos, cerca de 0,2, ainda são baixos, o que sugere que o *k-means* não conseguiu lidar com a alta dimensionalidade dos dados criando grupos com baixa coesão.

4.3 Algoritmo *Canopy*

Os resultados obtidos por meio do agrupamento *canopy*, tiveram diferença tanto nas bases de doenças (Tabelas 8 e 9) quanto nas bases de acidentes (Tabela 10 e 11).

Nas bases geradas a partir da codificação inteira-binária houve um pequeno aumento no índice da silhueta simplificada à medida que o número de grupos aumenta. Porém as bases geradas pela codificação 1-de-n, começaram com um valor de silhueta mais alto e a medida que foi aumentando o número de grupos, foi diminuindo esses valores, ou seja, os melhores resultados foram para o valor de agrupamento bem baixo. Isto pode indicar que estas bases não possuem uma estrutura de grupo ou que o algoritmo *canopy* não conseguiu identificar grupos nestas bases.

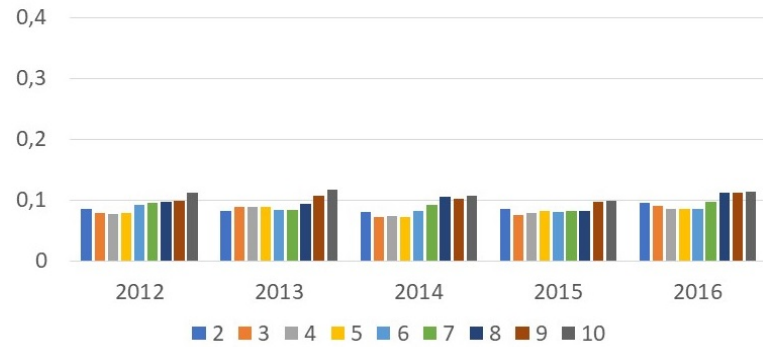
A partir das Figuras 15 e 16 pode-se ter uma visão melhor da diferença entre os resultados das bases.

Tabela 8 – Agrupamento *canopy* - Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.

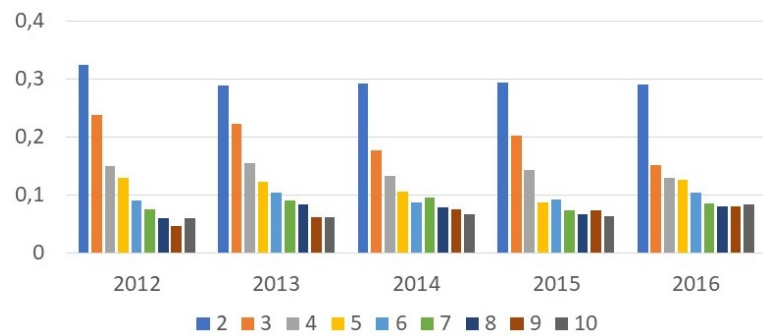
Nº de agrupamento	2012	2013	2014	2015	2016
2	0,085129126	0,082579012	0,081058011	0,084852331	0,095564376
3	0,078779136	0,088315179	0,072299217	0,07569551	0,089656141
4	0,076835152	0,089044447	0,073521931	0,079307349	0,085655991
5	0,078983275	0,089136335	0,07200748	0,082298653	0,085421334
6	0,091506005	0,082967593	0,08225935	0,080621844	0,086103322
7	0,095589298	0,083314838	0,092766769	0,081202926	0,097443589
8	0,097233408	0,094262851	0,10519008	0,082285544	0,111293536
9	0,098681127	0,107531265	0,101694703	0,096605622	0,112404713
10	0,111627049	0,117215773	0,106607818	0,098117271	0,113060219

Tabela 9 – Agrupamento *canopy* - Base_ano_1n_D, onde o ano vai de 2012 à 2016.

Nº de agrupamento	2012	2013	2014	2015	2016
2	0,323959004	0,288196254	0,292616277	0,293727323	0,290523333
3	0,237180092	0,221786787	0,177394421	0,201815414	0,152246547
4	0,149417242	0,154360781	0,133105459	0,142885682	0,129900476
5	0,129593674	0,12321708	0,106316365	0,087548521	0,126356649
6	0,091190473	0,1035589	0,08742219	0,092985774	0,10366587
7	0,075574135	0,090791726	0,094709026	0,073767711	0,085278131
8	0,059192914	0,083237373	0,078207155	0,066892496	0,080910609
9	0,04683049	0,061875694	0,074408339	0,073150039	0,079891823
10	0,060383284	0,062277091	0,066542613	0,063818928	0,083394181



(a) Agrupamento *canopy* - Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.



(b) Agrupamento *canopy* - Base_ano_1n_D, onde o ano vai de 2012 à 2016.

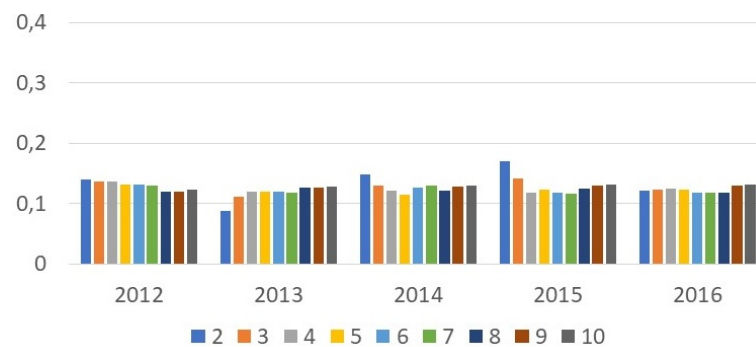
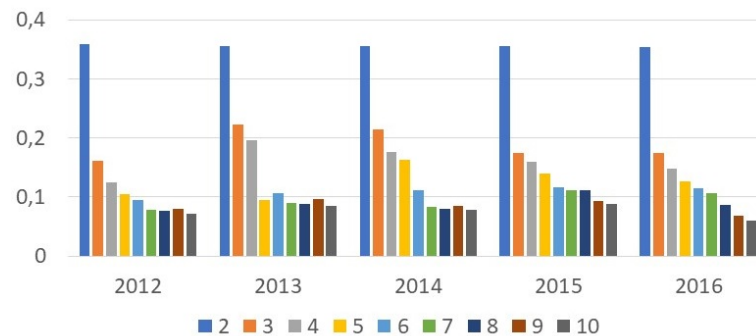
Figura 15 – Gráficos das bases de dados de doenças utilizando o algoritmo *canopy* e modificando o número de agrupamento.

Tabela 10 – Agrupamento *canopy* - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.

Nº de agrupamento	2012	2013	2014	2015	2016
2	0,139356715	0,0882482	0,147590383	0,170527936	0,122107085
3	0,137161806	0,111259466	0,130629892	0,140752652	0,122976843
4	0,136616439	0,118914596	0,121140603	0,117764064	0,124888512
5	0,13148902	0,119551415	0,113876426	0,122756612	0,123240737
6	0,131188021	0,120382476	0,126296033	0,11749506	0,117432696
7	0,129058625	0,117368784	0,129333598	0,117022148	0,117201797
8	0,120255582	0,126701831	0,121411145	0,124991017	0,118006633
9	0,120173886	0,12717468	0,128548557	0,129819855	0,130110314
10	0,123362692	0,127876078	0,129589333	0,131308222	0,131257026

Tabela 11 – Agrupamento *canopy* - Base_ano_1n_A, onde o ano vai de 2012 à 2016.

Nº de agrupamento	2012	2013	2014	2015	2016
2	0,358806058	0,355694349	0,355306587	0,356162991	0,354830142
3	0,160702263	0,222925571	0,213996959	0,175289475	0,175205612
4	0,124435907	0,196319096	0,175809562	0,159704438	0,148275993
5	0,10522472	0,095333151	0,163278197	0,139331235	0,126069894
6	0,095498656	0,106592119	0,111436033	0,115741287	0,114820507
7	0,078056441	0,09063343	0,083010403	0,110856541	0,106404148
8	0,075958016	0,087675981	0,080335701	0,110869718	0,086062033
9	0,079423121	0,096225443	0,084508378	0,093819285	0,068487291
10	0,071921965	0,085237942	0,07817382	0,088061539	0,059542214

(a) Agrupamento *canopy* - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.(b) Agrupamento *canopy* - Base_ano_1n_A, onde o ano vai de 2012 à 2016.Figura 16 – Gráficos das bases de dados de acidentes utilizando o algoritmo *canopy* e modificando o número de agrupamento.

Para valores pequenos de k , pode-se perceber que o *canopy* obtém resultados melhores que o *k-means*.

4.4 Algoritmo EM

O último algoritmo de agrupamento aplicado nas bases foi o EM. Os resultados deste algoritmo aplicado às bases geradas a partir da codificação inteira-binária, mostra que obteve valores baixos de silhueta e que não há grandes variações desses valores à medida que o valor de k aumenta.

O algoritmo EM apresentou piores resultados que o *k-means* e o *canopy* nas bases de dados geradas a partir da codificação 1-de-n. Em geral, muitos valores da silhueta ficaram próximos de zero e alguns ficaram negativos. Isso não é desejável, pois significa que a distância média dos objetos para o centroide do seu próprio grupo é maior que a distância média para os centroides dos outros grupos.

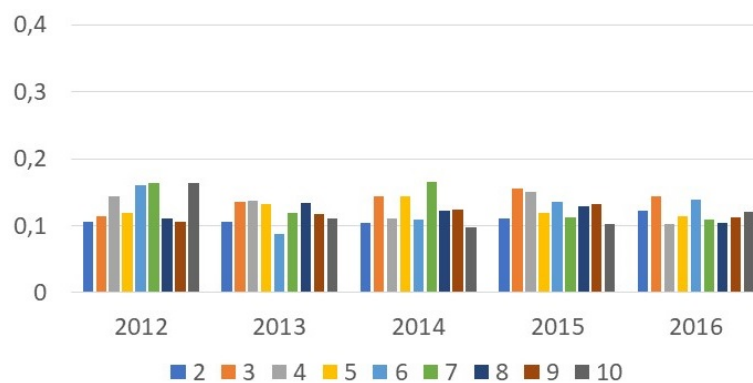
As bases de dados de doenças são apresentadas pelas Tabelas 12 e 13 e ilustrada pela Figura 17, já as bases de acidentes são representadas pelas Tabelas 14 e 15 e pela Figura 18.

Tabela 12 – Agrupamento EM - Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.

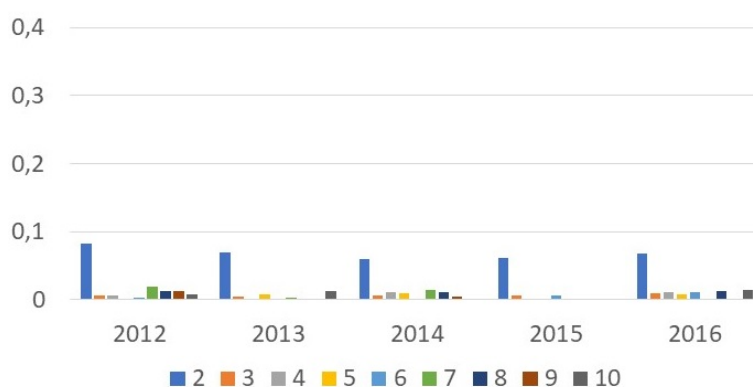
Nº de agrupamento	2012	2013	2014	2015	2016
2	0,10621788	0,105683444	0,1050816	0,111174963	0,123155049
3	0,114682754	0,135043852	0,143971841	0,156086608	0,144596862
4	0,14361561	0,13739291	0,111638035	0,150647053	0,10196185
5	0,119435552	0,132999797	0,143237697	0,119169343	0,114252428
6	0,1603817	0,087282202	0,109020146	0,135076281	0,138989081
7	0,164413349	0,118905506	0,165107102	0,113343101	0,108848891
8	0,110360328	0,133931334	0,122727875	0,129458038	0,103908095
9	0,106716127	0,118310351	0,123528405	0,131911635	0,111909741
10	0,163794635	0,110710175	0,097779238	0,102703841	0,121612673

Tabela 13 – Agrupamento EM - Base_ano_1n_D, onde o ano vai de 2012 à 2016.

Nº de agrupamento	2012	2013	2014	2015	2016
2	0,081434099	0,069176238	0,058937882	0,060376441	0,067948728
3	0,005897295	0,003927996	0,005646779	0,006568631	0,008945325
4	0,005284851	-0,003859849	0,011284319	-0,000239558	0,010489644
5	-0,001914848	0,00725892	0,009869519	-0,008187948	0,006803674
6	0,002838338	0,001235517	0,001422865	0,006355336	0,009982816
7	0,018593077	0,003280584	0,013696558	0,001453638	0,000110556
8	0,012735868	-0,010798323	0,010172476	-0,017855957	0,011932766
9	0,012064324	-0,019076029	0,003833486	-0,01935109	0,000223644
10	0,007111826	0,012164958	-0,018888268	-0,016557245	0,014225254



(a) Agrupamento EM, Base_ano_IntBin_D, onde o ano vai de 2012 à 2016.



(b) Agrupamento EM, Base_ano_1n_D, onde o ano vai de 2012 à 2016.

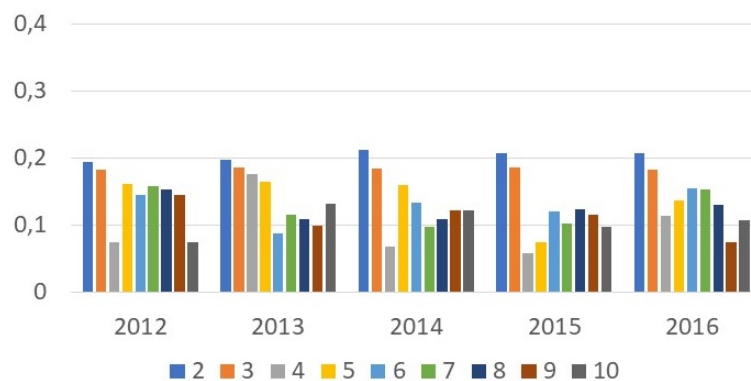
Figura 17 – Gráficos das bases de dados de doenças utilizando o algoritmo EM e modificando o número de agrupamento.

Tabela 14 – Agrupamento EM - Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.

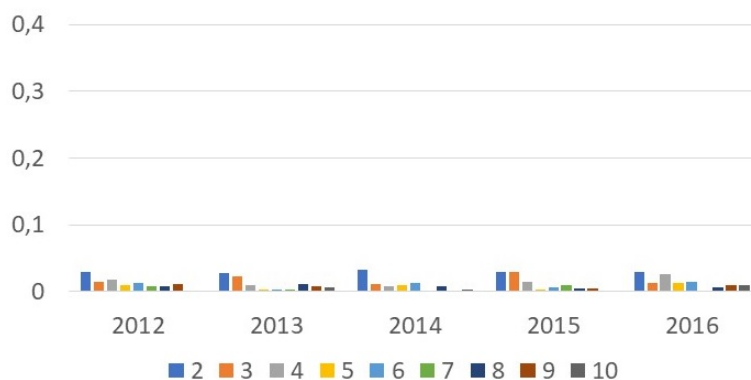
Nº de agrupamento	2012	2013	2014	2015	2016
2	0,193806166	0,19715182	0,211258498	0,206891553	0,206342629
3	0,182668036	0,185030763	0,184309242	0,185887299	0,182710488
4	0,075147272	0,17599917	0,068613232	0,057821146	0,113522398
5	0,160571219	0,164326031	0,159964598	0,074553406	0,136367071
6	0,144753338	0,087668368	0,133557212	0,120911648	0,154323327
7	0,157160982	0,116169118	0,096936305	0,101648646	0,152573483
8	0,153481026	0,108242879	0,109266202	0,123933343	0,130332575
9	0,144367457	0,098638335	0,121315698	0,115119098	0,07518105
10	0,074876078	0,13178701	0,121913771	0,097983813	0,107870651

Tabela 15 – Agrupamento EM - Base_ano_1n_A, onde o ano vai de 2012 à 2016.

Nº de agrupamento	2012	2013	2014	2015	2016
2	0,029027831	0,027314946	0,032405849	0,029977769	0,030001361
3	0,01376888	0,023566561	0,010516673	0,029177606	0,012319073
4	0,018521366	0,010203176	0,008297201	0,014040875	0,025532487
5	0,010344984	0,002230649	0,010306084	0,003553477	0,013365189
6	0,012036018	0,003698771	0,013210112	0,005810401	0,013972484
7	0,007223917	0,002514373	-0,001222205	0,009416638	-0,001446366
8	0,008101122	0,011401306	0,008320379	0,004574362	0,005895105
9	0,011298314	0,007443061	-0,00351939	0,004639546	0,00901073
10	-0,147601172	0,005794656	0,003774298	-0,008368341	0,009974936



(a) Agrupamento EM, Base_ano_IntBin_A, onde o ano vai de 2012 à 2016.



(b) Agrupamento EM, Base_ano_1n_A, onde o ano vai de 2012 à 2016.

Figura 18 – Gráficos das bases de dados de acidentes utilizando o algoritmo EM e modificando o número de agrupamento.

4.5 Considerações Finais

Este capítulo apresentou os resultados obtidos pelos algoritmos *k-means*, *canopy* e EM no agrupamento da base de dados de benefícios previdenciários. Pode-se perceber que

os algoritmos utilizados não trouxeram bons resultados para o agrupamento das bases de dados, pois os valores de silhueta foram baixos. Uma das possíveis motivações para tal comportamento é que as bases de dados possuem alta dimensionalidade após terem sido pré-processadas e tais algoritmos possuem dificuldade em trabalhar com alta dimensionalidade.

Também pode-se perceber que variações nos valores de k não alteraram o comportamento dos algoritmos. Embora, ainda seja necessário testar valores mais altos de k .

Essa foi uma investigação inicial e notou-se que os primeiros passos não conduziram a bons resultados, isto merece uma investigação mais profunda. O foco deste trabalho era ter um pré-processamento inicial da base, fazer alguns experimentos iniciais e a partir disso ter um indicativo de quais caminhos devem ou não ser melhor investigados.

O próximo capítulo irá relatar as principais contribuições deste trabalho e ideias para possíveis trabalhos futuros.

5 Conclusão

Este trabalho apresentou algumas técnicas de mineração de dados aplicadas nos dados de benefícios previdenciários do Ministério Público do Trabalho. Foi criada uma ferramenta na linguagem Java capaz de realizar algumas técnicas de pré-processamento como: remover dados, converter alguns atributos para uma versão mais resumida, converter dados categóricos para numéricos, realizar codificações inteira-binária e 1-de-n e reescalar os dados.

Após realizar a etapa de pré-processamento nos dados, foi criada diferentes bases afim de explorar técnicas de agrupamento em diferentes visões. As técnicas de agrupamento utilizadas foram o *k-means*, *canopy* e EM, sendo alterado em cada execução, a quantidade de grupos formados. Para validar esses agrupamentos foi utilizado o índice de Silhueta Simplificada, que é uma técnica de validação de agrupamento com um bom desempenho computacional.

Por fim, após validar os agrupamentos concluiu-se que, com as técnicas de agrupamento utilizadas não foi possível encontrar padrões nos dados, visto que, os valores do índice de silhueta em todas as bases foram baixos. Com isso, foi possível ter um indicativo de quais caminhos não deve ser continuados a serem investigados.

5.1 Contribuições

Este trabalho proporcionou as seguintes contribuições para os dados do MPT:

- Criação de uma ferramenta em Java capaz de realizar etapas de pré-processamento na base de dados do MPT. Ainda que a ferramenta contenha alguns métodos específicos para a conversão dos dados do MPT, vários dos métodos de pré-processamento implementados podem ser usados em diferentes problemas;
- Aplicação de diferentes técnicas de agrupamento variando a quantidade de grupos;
- Avaliação dos resultados dos agrupamentos por meio do índice de Silhueta Simplificada.

5.2 Trabalhos Futuros

A partir dos resultados, é possível indicar possíveis trabalhos futuros para melhorar o agrupamento dos dados do MPT:

- Utilizar métodos de agrupamento mais complexos e que explorem outras vertentes, tais como: agrupamento relacional, agrupamento por densidade, agrupamento em grid e hierárquico;
- Utilizar diferentes medidas de validação de agrupamento;
- Aplicar técnicas de visualização nos resultados obtidos;
- Aplicar técnicas de seleção de atributos a fim de reduzir a dimensionalidade das bases pré-processadas;
- Investigar outras técnicas de pré-processamento dos dados;
- Continuar a investigação na busca de padrões na base de dados de benefícios que possam auxiliar o trabalho de gestores e especialistas em trabalho;
- Aplicar as mesmas técnicas de pré-processamento e agrupamento de dados na base de CATs.

Referências

- ADMMC. *Acidentes de trabalho no Brasil: como diminuir os números alarmantes* ? 2016. Disponível em: <<https://www.mobussconstrucao.com.br/blog/2016/03/acidentes-de-trabalho-no-brasil/>>. Acesso em: 15 fev. 2018. Citado na página 14.
- ALVARES, G.; COSTA, I. Os impactos dos acidentes de trabalho. UBÁ, MG, 2015. Disponível em: <<https://www.webartigos.com/storage/app/uploads/public/588/508/229/588508229b3e5486118382.pdf>>. Citado na página 15.
- ALVES, V. S. et al. *Um algoritmo evolutivo rápido para agrupamento de dados*. Dissertação (Mestrado), 2007. Citado na página 35.
- AMO, S. de; ROC, C. *Curso de Data Mining*. Dissertação (Mestrado) — Universidade Federal de Uberlândia, 2003. Disponível em: <http://www.fatecead.com.br/tei/semana08-1_livro_mineracaodados.pdf>. Citado na página 26.
- ARAÚJO, J. P. *Manual dos direitos dos segurados do INSS*. 2011. Disponível em: <http://www.graficosantista.org.br/e107_files/downloads/cartilha_inss.pdf>. Acesso em: 25 junho 2018. Citado na página 17.
- BAOHUA, G.; FEIFANG, H.; HUAN, L. *Sampling and its application in data mining: A survey*. [S.l.], 2000. Disponível em: <<http://dl.comp.nus.edu.sg/bitstream/handle/1900.100/1408/report.pdf?sequence=3&isAllowed=y>>. Citado na página 23.
- BARNETT, V.; LEWIS, T. *Outliers in statistical data*. [S.l.]: Wiley, 1974. Citado na página 22.
- BARTOLOMEU, T. A. et al. *Modelo de investigação de acidentes do trabalho baseado na aplicação de tecnologias de extração de conhecimento*. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2002. Disponível em: <<https://repositorio.ufsc.br/bitstream/handle/123456789/83836/189111.pdf?sequence=1>>. Citado 3 vezes nas páginas 13, 15 e 35.
- BORGES, V. R. P. Comparação entre as técnicas de agrupamento k-means e fuzzy c-means para segmentação de imagens coloridas. 2010. Citado na página 46.
- BRASIL. *Anuário Estatístico da Previdência Social*. 2016. Disponível em: <<http://www.previdencia.gov.br/wp-content/uploads/2018/01/AEPS-2016.pdf>>. Acesso em: 16 fev. 2018. Citado 2 vezes nas páginas 15 e 16.
- BRASIL. *Previdência Social. CNP: Conselho de Previdência discute Saúde e Segurança no Trabalho*. 2017. Disponível em: <<http://www.previdencia.gov.br/2017/05/cnp-conselho-de-previdencia-discute-saude-e-seguranca-no-trabalho>>. Acesso em: 10 set. 2017. Citado na página 10.
- CAETANO, A. J.; MACHADO, C. J. Consistência e identificabilidade no modelo grade of membership: uma nota metodológica. *Revista Brasileira de Estudos de População*, SciELO Brasil, v. 26, n. 1, p. 145–149, 2009. Citado na página 30.

- CAMPELLO, R. J. G. B. *Análise de Agrupamento de Dados*. 2014. Disponível em: <http://wiki.icmc.usp.br/images/2/21/Algoritmos_Particionais_II.pdf>. Acesso em: 05 julho 2018. Citado 2 vezes nas páginas 30 e 31.
- CASSIANO, K. M. *Análise de Séries Temporais Usando Análise Espectral Singular (SSA) e Clusterização de Suas Componentes Baseada em Densidade*. Tese (Doutorado) — PUC-Rio, 2014. Disponível em: <http://www2.dbd.puc-rio.br/pergamum/tesesabertas/1012121_2014_pretextual.pdf>. Citado 2 vezes nas páginas 26 e 31.
- CASTANHEIRA, L. G. *Aplicação de técnicas de mineração de dados em problemas de classificação de padrões*. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, 2008. Disponível em: <https://www.ppgee.ufmg.br/documentos/Defesas/777/Dissertacao_LucianaCastanheira.pdf>. Citado 3 vezes nas páginas 21, 25 e 34.
- CIPA. C.A.T. - *Comunicação de Acidente no Trabalho*. 2011. Disponível em: <<http://cipa.fmrp.usp.br/Html/CAT.htm>>. Acesso em: 06 dez. 2017. Citado na página 15.
- CONCEIÇÃO, V. S. et al. Desenvolvimento de critérios de validação de múltiplos agrupamentos em mapreduce. 2015. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/eniac/2015/009.pdf>>. Citado 2 vezes nas páginas 35 e 46.
- COSTA, M. da. *Cartilha de direito previdenciário*. 2017. Disponível em: <<http://www.jurisite.com.br/wordpress/wp-content/uploads/2017/03/CARTILHA-PREVIDENCI%C3%81RIA.pdf>>. Acesso em: 25 junho 2018. Citado 2 vezes nas páginas 16 e 17.
- DATASUS. *Mesorregião*. 2018. Disponível em: <<http://datasus.saude.gov.br/cadastros-nacionais/309-mesorregiao>>. Acesso em: 28 abril. 2018. Citado na página 42.
- DONI, M. V. Análise de cluster: métodos hierárquicos e de particionamento. *Universidade Presbiteriana Mackenzie*, 2004. Disponível em: <<http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF>>. Citado 2 vezes nas páginas 26 e 27.
- DUARTE, J. C. *Auxílio-doença acidentário, auxílio-doença e auxílio-acidente*. 2015. Disponível em: <<https://josianeclemente.jusbrasil.com.br/artigos/207012251/auxilio-doenca-acidentario-auxilio-doenca-e-auxilio-acidente>>. Acesso em: 02 julho 2018. Citado na página 10.
- EDUCAÇÃO, S. da. *Mesorregiões Geográficas*. 2010. Disponível em: <<http://www.geografia.seed.pr.gov.br/modules/galeria/detalhe.php?foto=1586&evento=8>>. Acesso em: 28 abril. 2018. Citado na página 42.
- EIBE, F.; MARK, A. H.; IAN, H. W. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition*. 2016. Citado na página 46.
- ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. [s.n.], 1996. v. 96, n. 34, p. 226–231. Disponível em: <<http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf>>. Citado 2 vezes nas páginas 31 e 32.

- FABRICIO, J. G. e. H. M. G. J. A. *Descoberta de conhecimento utilizando o processo KDD*. 2017. Disponível em: <<https://www.devmedia.com.br/descoberta-de-conhecimento-utilizando-o-processo-kdd/38709#>>. Acesso em: 16 nov. 2017. Citado na página 20.
- FAYYAD et al. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, 1996. Citado na página 19.
- FERREIRA, A. L. *Mapas auto organizáveis na descoberta e validação de padrões em bases de dados*. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2012. Citado na página 46.
- GEP/MTSS. Introdução à saúde e segurança no trabalho. 2009. Disponível em: <http://www.ilo.org/public/portugue/region/eurpro/lisbon/pdf/pub_modulos2.pdf>. Acesso em: 03 nov. 2017. Citado na página 10.
- GOLDSCHMIDT, R.; PASSOS, E. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. *Rio de Janeiro: Campus*, v. 1, 2005. Citado 2 vezes nas páginas 19 e 21.
- GOLDSCHMIDT, R.; PASSOS, E. *Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações*. [S.l.]: Elsevier Brasil, 2017. Citado na página 20.
- GROSS, J. L. G. *URSA: um framework para agrupamento de dados e validação de resultados*. 2014. Monografia (Bacharel em Ciência da Computação), Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: <<https://www.lume.ufrgs.br/bitstream/handle/10183/110328/000952575.pdf?sequence=1>>. Citado 2 vezes nas páginas 28 e 34.
- GUIMARÃES, W. S. A. et al. *Data mining aplicado ao serviço público, extração de conhecimento das ações do Ministério Público Brasileiro*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2000. Disponível em: <<https://repositorio.ufsc.br/bitstream/handle/123456789/78868/194202.pdf?sequence=1>>. Citado na página 36.
- HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman Editora, 2009. Nenhuma citação no texto.
- HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. 83-445 p. Citado 3 vezes nas páginas 21, 22 e 26.
- HAND, D. J. Principles of data mining. *Drug safety*, Springer, v. 30, n. 7, p. 621–622, 2007. Citado na página 12.
- HO, R. *Pragmatic Programming Techniques*. 2011. Disponível em: <<http://horicky.blogspot.com/2011/04/k-means-clustering-in-map-reduce.html>>. Acesso em: 16 junho 2018. Citado na página 30.
- 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. Citado na página 34.
- HUI, J. *Machine learning - Clustering, Density based clustering and SOM*. 2017. Disponível em: <<https://jhui.github.io/2017/01/15/Machine-learning-clustering/>>. Acesso em: 16 junho 2018. Citado na página 30.

- IBGE. *CONCLA Comissão Nacional de Classificação*. 2018. Disponível em: <<https://cnae.ibge.gov.br/?view=estrutura>>. Acesso em: 22 maio 2018. Citado na página 41.
- ICLINIC. *O que é CID 10? Saiba mais sobre seus principais códigos!* 2016. Disponível em: <<https://blog.iclinic.com.br/o-que-e-cid-10/>>. Acesso em: 21 abril. 2018. Citado na página 38.
- INSS, A. de Comunicação Social do. *Comunicação de Acidente de Trabalho - CAT*. 2017. Disponível em: <<https://portal.inss.gov.br/servicos-do-inss/comunicacao-de-acidente-de-trabalho-cat>>. Acesso em: 11 set. 2017. Citado 2 vezes nas páginas 10 e 15.
- JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010. Citado na página 27.
- JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. [S.l.]: Prentice-Hall, Inc., 1988. Citado na página 34.
- KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial intelligence*, Elsevier, v. 97, n. 1-2, p. 273–324, 1997. Citado na página 23.
- KRUMHEUER, E. *CNAE: O que é? Aprenda a classificar sua empresa*. 2017. Disponível em: <<https://blog.contaazul.com/cnae-o-que-e-confira-tabela/>>. Acesso em: 22 abril. 2018. Citado na página 38.
- LAROSE, D. T. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2005. 31 p. Citado 2 vezes nas páginas 22 e 26.
- LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, v. 4, p. 18–36, 2009. Citado na página 11.
- MADEIRA, F. D. *Mineração de Dados Educacionais usando KDD - Parte 1*. 2013. Disponível em: <<https://www.devmedia.com.br/mineracao-de-dados-educacionais-usando-kdd-parte-1/28968#>>. Acesso em: 12 set. 2017. Citado na página 11.
- MAHOUT. *Canopy Clustering*. 2017. Disponível em: <<https://mahout.apache.org/users/clustering/canopy-clustering.html>>. Acesso em: 16 junho 2018. Citado na página 30.
- MANNILA, H. Data mining: machine learning, statistics, and databases. In: IEEE. *Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on*. [S.l.], 1996. p. 2–9. Citado na página 21.
- MENDES, A. d. S. *Aplicação de técnicas de data mining na caracterização de turnover interno para o suporte à gestão de pessoas*. Dissertação (Mestrado) — Universidade de Brasília, 2014. Citado na página 24.
- MENDES, L. *Data Mining - Estudo de Técnicas e Aplicações na Área Bancária*. 2011. Monografia (Tecnólogo em Processamento de Dados), Faculdade de Tecnologia de São Paulo. Citado na página 24.

- MENDONÇA, L. O. *Abril Verde: mês dedicado à prevenção de acidentes de trabalho e doenças ocupacionais*. 2017. Disponível em: <<http://justificando.cartacapital.com.br/2017/04/12/abril-verde-mes-dedicado-prevencao-de-acidentes-de-trabalho-e-doencas-ocupacionais>>. Acesso em: 10 set. 2017. Citado na página 10.
- MERELES, C. *Ministério Público do Trabalho: qual a sua função?* 2017. Disponível em: <<http://www.politize.com.br/ministerio-publico-do-trabalho>>. Acesso em: 11 set. 2017. Citado na página 10.
- MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. C. Machine learning, neural and statistical classification. CiteSeer, 1994. Citado na página 25.
- MIRANDA, L. B. A. de. *Análise de Clusters com Restrições de Contiguidade Espacial*. Dissertação (Mestrado) — Universidade Federal Fluminense, 2016. Citado na página 29.
- MORAES, G. *Legislação de Segurança e Saúde no trabalho: normas regulamentadoras do Ministério do Trabalho e Emprego*. [S.l.]: Virtual, 2012. Citado na página 16.
- MPT. *O Ministério Público do Trabalho e as questões de segurança e saúde no trabalho*. 2013. Disponível em: <http://www.pgt.mpt.gov.br/publicacoes/seguranca/mpt_questoes_seg_saude.pdf>. Acesso em: 07 dez. 2017. Citado na página 17.
- MPT. *O Ministério Público do Trabalho e os Direitos dos Trabalhadores*. 2014. Disponível em: <<http://www.pcdlegal.com.br/cartilhampt/dvisual/capitulo18.php>>. Acesso em: 06 dez. 2017. Citado na página 14.
- MPT. *Ministério Público do Trabalho*. 2015. Disponível em: <http://portal.mpt.mp.br/wps/portal/portal_mpt/mpt/ompt/mpt/>. Acesso em: 06 dez. 2017. Citado na página 17.
- NASSIF, L. F. d. C. *Técnicas de agrupamento de textos aplicadas à computação forense*. Dissertação (Mestrado) — Universidade de Brasília, 2011. Citado na página 46.
- OBSERVATÓRIO. *Observatório Digital de Saúde e Segurança no Trabalho - Smartlab de Trabalho Decente MPT - OIT*. 2017. Disponível em: <<http://observatoriosst.mpt.mp.br>>. Acesso em: 13 set. 2017. Citado 6 vezes nas páginas 11, 12, 18, 19, 35 e 36.
- OIT. *MPT e OIT lançam Observatório Digital de Saúde e Segurança do Trabalho*. 2017. Disponível em: <http://www.ilo.org/brasilia/noticias/WCMS_551769/lang-pt/index.htm>. Acesso em: 24 fev. 2018. Citado na página 18.
- OLIVEIRA, R. A. d. *Algoritmos para determinação do número de grupos em estudos de formas planas*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2016. Disponível em: <https://repositorio.ufpe.br/bitstream/handle/123456789/17314/Rodrigo_CD.pdf?sequence=1&isAllowed=y>. Citado na página 28.
- PANDRE, A. *Cluster Analysis: see it 1st*. 2009. Disponível em: <<https://apandre.wordpress.com/visible-data/cluster-analysis/>>. Acesso em: 07 dez. 2017. Citado na página 29.
- PANTALEÃO, S. *Emissão da CAT no contrato temporário ou de experiência gera estabilidade ?* 2016. Disponível em: <http://www.guiatrabalhista.com.br/tematicas/cat_temporario.htm>. Acesso em: 21 fev. 2018. Citado na página 15.

- PIGNATA, A. R. *Clusterização de dados sobre Comunicação de Acidente de Trabalho*. [S.l.]: Brasília, 2016. Paper (Seminário em Ciência Política: Mineração de dados), Instituto de Ciência Política da Universidade de Brasília. Citado 2 vezes nas páginas 13 e 36.
- PJERJ. *Benefícios em Espécie*. 2014. Disponível em: <<http://www.tjrj.jus.br/documents/10136/110217/beneficios-em-especie.pdf>>. Acesso em: 05 julho 2018. Citado na página 16.
- PREVIDÊNCIA. *RGPS: Déficit da Previdência Social em 2016 foi de R\$ 151,9 bilhões*. 2017. Disponível em: <<http://www.previdencia.gov.br/2017/01/rgps-deficit-da-previdencia-social-em-2016-foi-de-r-1519-bilhoes/>>. Acesso em: 21 fev. 2018. Citado na página 15.
- REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003. Citado na página 21.
- SCHMITT, J. et al. Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo. Florianópolis, SC, 2005. Citado na página 12.
- SEMAAN, G. S. *Algoritmos para o Problema de Agrupamento Automático*. Tese (Doutorado) — Tese de Doutorado, Instituto de Computação, Universidade Federal Fluminense, 2013. Citado 2 vezes nas páginas 32 e 33.
- SINCOVAGA, R. *Entenda o que é e como funciona o acidente de trabalho e a CAT*. 2012. Disponível em: <<http://sincovaga.com.br/dl/comunicados/ENTENDA%20O%20QUE%20%C3%89%20E%20COMO%20FUNCIONA%20O%20ACIDENTE%20DE%20TRABALHO%20E.pdf>>. Acesso em: 05 dez. 2017. Citado na página 14.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Ciência Moderna, 2009. Citado 14 vezes nas páginas 20, 21, 22, 23, 24, 25, 26, 28, 31, 32, 33, 34, 40 e 41.
- VALADA, M. M. *Fiscalização em Segurança e Saúde no Trabalho*. 2015. Disponível em: <<https://matheusvalada.jusbrasil.com.br/artigos/178744392/fiscalizacao-em-seguranca-e-saude-no-trabalho>>. Acesso em: 31 maio 2018. Citado na página 17.
- VALE, M. N. do. *Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos*. Tese (Doutorado) — PUC-Rio, 2005. Citado na página 27.
- VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. *Stat. Anal. Data Min.*, John Wiley & Sons, Inc., New York, NY, USA, v. 3, n. 4, p. 209–235, ago. 2010. ISSN 1932-1864. Disponível em: <<http://dx.doi.org/10.1002/sam.v3:4>>. Citado na página 35.
- WAIKATO, T. U. of. *Attribute-Relation File Format (ARFF)*. 2008. Disponível em: <<https://www.cs.waikato.ac.nz/ml/weka/arff.html>>. Acesso em: 29 maio 2018. Citado na página 43.
- WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2011. Citado 2 vezes nas páginas 22 e 46.