UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DEPARTAMENTO ACADÊMICO DE INFORMÁTICA

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS RAFAEL ANDRADE

USO DE MINERAÇÃO DE DADOS PARA DESCOBERTA DE REGRAS DE ASSOCIAÇÃO EM PRONTUÁRIOS MÉDICOS

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA

2019

LUCAS RAFAEL ANDRADE

USO DE MINERAÇÃO DE DADOS PARA DESCOBERTA DE REGRAS DE ASSOCIAÇÃO EM PRONTUÁRIOS MÉDICOS

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação, do Departamento de Informática, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. André Pinz Borges

PONTA GROSSA

2019

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Ponta Grossa

Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática
Bacharelado em Ciência da Computação

TERMO DE APROVAÇÃO

USO DE MINERAÇÃO DE DADOS PARA DESCOBERTA DE REGRAS DE ASSOCIAÇÃO NA ÁREA DA SAÚDE

por

LUCAS RAFAEL ANDRADE

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 11 de novembro de 2019 como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof. Dr. André Pinz Borges
Orientador(a)

_____
Profa Dra. Helyane Bronoski Borges
Membro titular

_____
Prof. Dr. Richardson Ribeiro (UFPR)
Membro titular

_____
Prof. MSc Geraldo Ranthum
Responsável pelo Trabalho de Conclusão de Curso

_____
Profa. Dra. Mauren Louise Sguario
Coordenador do curso

Dedico este trabalho à minha família e amigos.

AGRADECIMENTOS

Agradeço à minha mãe Sara e meu pai Marcos, que estiveram sempre junto comigo fornecendo todo o apoio do mundo nesta jornada, nos momentos bons e nos momentos difíceis.

Ao meu professor e orientador Dr. André Pinz Borges com quem aprendi muito seja através de suas aulas, orientações, revisões e reuniões.

Aos colegas de faculdade e amigos Filipe, Kaique, Fernando, Tiago, Nataly, que também estiveram batalhando neste caminho.

Aos meus amigos Caio, Fellipe, Guilherme, Roberto e Wesley com quem mantive contato por quase toda a minha jornada acadêmica e que foram parte essencial no meu desenvolvimento intelectual e cultural.

A toda a literatura, cinema, música, jogos e artes que me mantiveram sempre inspirado e que também, de certa forma, me acompanham durante esses anos.

Agradeço a todos que não foram citados aqui mas que também foram de alguma forma responsáveis pela realização deste trabalho.

Qualquer tecnologia suficientemente
avançada é indistinguível da magia.
(CLARKE, Arthur C., 1962)

RESUMO

ANDRADE, Lucas Rafael. **Uso de mineração de dados para descoberta de regras de associação em prontuários médicos.** 85 páginas. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2019.

Este trabalho tem como objetivo a extração de regras de associação através de técnicas de mineração de dados em uma base de dados de prontuários eletrônicos. As regras de associação tornam-se, após análise, informação útil para o profissional da área da saúde. O processo de descoberta de informações em prontuários pode ser demorado e complexo para o profissional sem o auxílio da tecnologia especializada para o processo. Para tal, a execução dos processos de descoberta de conhecimentos em bases de dados se mostra uma opção viável para adquirir informações e conhecimento útil. A base de dados utilizada neste trabalho compreende prontuários eletrônicos referentes a 43.879 pacientes usuários do sistema único de saúde e um total de 2.296.626 atendimentos realizados no ano de 2015 na cidade de Pato Branco, Paraná. Para realização do processo de preparação dos dados foi utilizado o *software* PostgreSQL, onde é possível manipular e realizar alterações ao banco de dados e para a aplicação dos algoritmos *Apriori* e *Hotspot* na etapa de Mineração de Dados foi utilizado o WEKA. Realizou-se experimentos com a aplicação de Apriori onde compara-se a discrepância de regras usando a métrica de confiança com a utilização da métrica lift. Os resultados do experimento evidenciaram regras de associação que não seriam visualizadas através da execução do algoritmo considerando somente a métrica de confiança. Com os resultados dos experimentos, foi inspirada uma escolha de atributos para a execução do algoritmo Hotspot, onde, primeiramente, foi observada as diferenças de doenças entre o sexo masculino e feminino. Também foi traçado o perfil de cada uma das faixas etárias presentes na base de dados, onde foram encontradas doenças que não apareceriam numa execução do Apriori, especialmente em faixas de 18 anos para cima. Foi realizada a execução do algoritmo Hotspot também para três grupos de doenças que apareceram dentre as faixas etárias. Os grupos estavam relacionados a dorsopatias, hipertensão e doenças médias no ouvido e mastóide. Das regras obtidas, pôde-se fortalecer a relação dentre as doenças e as faixas etárias das quais elas foram baseadas como também encontrar outras faixas etárias onde ocorrências e correlacionar o grupo de doenças com outros sintomas, especificando o perfil traçado das doenças. Para a conclusão do processo de descoberta de conhecimentos, mostra-se necessário a análise de um profissional da área da saúde sobre as informações obtidas a partir da aplicação de mineração de dados.

Palavras-chave: Mineração de Dados. Saúde. Prontuário Eletrônico. Associação. Hotspot.

ABSTRACT

ANDRADE, Lucas Rafael. **Use of Data Mining for association rules discovery in medical records**. 85 pages. Work of Conclusion Course (Graduation in Computer Science) - Federal Technology University - Paraná. Ponta Grossa, 2019.

This study aims to extract association rules in a medical records database through data mining techniques. Association rules can become useful information for health professionals, after analysis. For professionals, the information discovery in medical records can be complex and long without specialized technology's aid on the process. For this purpose, the application of knowledge discovery in databases shows an viable option for acquiring useful knowledge and information. The database used in this study consists of medical records about 43.879 patients and 2.296.626 health care attendances, on the year of 2015, in Pato Branco city, Paraná. The PostgreSQL software was used for the purpose of data preparation, where you can alter the database, whereas WEKA was used for data mining application of the Apriori and Hotspot algorithms. Tests where the difference between the use of the confidence metric and the lift metric in the Apriori algorithm were made. The results showed association rules that wouldn't be seen by considering the confidence metric Apriori application only. Using the results, attributes were then chosen for the Hotspot algorithm application, were, first the difference between disease groups from the male and female genders. Profiles for the age gaps that appeared on the database were made and the application found disease groups that wouldn't appear on the Apriori application, such as age gaps from 18 years and above. The Hotspot application was then used for profiling three disease groups which appeared between the age gaps. The groups were related to back pain, ear and mastoid diseases and hypertension. From the association rules obtained, the relation between age gaps and disease groups was made clearer, with the rules even finding some different age gaps for diseases. The application also related the diseases with other symptoms, narrowing down the profiling for the diseases. The analysis of an health professional would be necessary for the conclusion of the study results, where the rules could then become useful knowledge.

Keywords: Data Mining. Health. Electronic Medical Record. Association. Hotspot.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

AB	Atenção Básica
ARFF	*Attribute Relation File Format*
CID-10	Classificação Internacional de Doenças - Décima Revisão
CPF	Cadastro de Pessoa Física
CSV	*Comma Separated Values*
DATASUS	Departamento de Informática do Sistema Único de Saúde
IOM	*Institute of Medicine*
KDD	*Knowledge Discovery in Databases*
MD	Mineração de Dados
PEC	Prontuário Eletrônico do Cidadão
PEP	Prontuário Eletrônico do Paciente
SOAP	*Simple Object Acess Protocol*
SUS	Sistema Único de Saúde
SQL	*Structured Querry Language*
WEKA	*Wakaito Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

O armazenamento de dados constitui uma forma importante de obtenção de informações nos dias de hoje, seja em bancos, lojas, lanchonetes, hospitais e em outros âmbitos. Diariamente surgem dados em grandes quantias e das mais variadas fontes, trazendo uma necessidade de desenvolvimento de ferramentas que possam desvendar e gerar conhecimentos a partir de tais bases (HAN, 2012). Um exemplo consiste no prontuário eletrônico, uma ferramenta capaz de guardar dados como sinais vitais, prescrições médicas, exames (solicitações e resultados), além de dados pessoais de pacientes como idade, nome, altura, doenças, profissão, entre outras (DATASUS, 2018).

Os prontuários eletrônicos são preenchidos por profissionais da área da saúde a cada atendimento à um paciente (DATASUS, 2018). A aquisição de dados é realizada através de consultas em hospitais e postos de saúde, a partir de consultas realizadas pelos pacientes, onde é possível formar um banco de dados com dados dos pacientes atendidos em tais estabelecimentos.

Iniciativas do Ministério de Saúde de digitalizar foram criadas, juntamente com a iniciativa do Sistema Único de Saúde (SUS), para a formação do DATASUS. O DATASUS é responsável por prover os órgãos do SUS sistemas de informação e auxílio de informática, nos projetos de planejamento, operação e controle (DATASUS, 2018).

O e-SUS AB (Atenção Básica) é uma estratégia, também do Ministério da Saúde, para reestruturar as informações de Atenção Básica em nível nacional, permitindo que a coleta de dados esteja inserida em atividades já desenvolvidas por profissionais. As possibilidades de utilização de tal estratégia estão adaptadas às realidades de cada município. Em Unidades Básicas de Saúde que contém computadores, é possível a utilização do sistema mesmo sem acesso à internet, fornecendo informações individuais (como idade, altura, dados pessoais, informações de morada, problemas de saúde, exames feitos) e funcionalidades que facilitam o dia a dia dos profissionais (PORTALDAB, 2018).

Uma vez que seja obtido o acesso aos dados, é possível a extração de informações não exploradas de uma base de dados, processo conhecido como Descoberta de Conhecimento em Bases de Dados (do inglês Knowledge Discovery in Databases, KDD). O foco do processo de descoberta do conhecimento se

entrelaça com diversos campos dentro da Inteligência Artificial, como Aprendizado de Máquinas e Redes Neurais (FAYYAD et al, 2013). Estão incluídos no processo de KDD cinco etapas: seleção, pré-processamento, mineração de dados, avaliação dos padrões e apresentação do conhecimento. Na etapa de seleção os dados são separados e selecionados em informações de interesse. Após o pré-processamento executado onde é aplicado, por exemplo, a transformação dos dados. Na sequência é aplicada a Mineração de Dados para descoberta de padrões, seguida pela interpretação ou avaliação dos padrões obtidos pelo algoritmo e apresentação do conhecimento (WITTEN et al, 2011).

A Mineração de Dados (MD) pode ser descrita como a procura e identificação de padrões úteis em dados. A partir da análise dos padrões pode ser gerado o conhecimento. A MD é um domínio que utiliza várias técnicas de áreas como Aprendizagem de Máquinas, Estatística, Matemática entre outras. Essa interdisciplinaridade contribui para o sucesso da MD e suas aplicações (HAN, 2012). Métodos de Aprendizagem de Máquina, Mineração de Dados e Regras de Associação são de interesse para aplicações na área da saúde pois apresentam um resultado em regras de se - então que estão mais alinhadas com o objetivo de criar padrões compreensíveis de modelos e conceitos médicos (KRYSTOF, 2002).

Dois tipos de algoritmos poderão ser usados para a MD, Regras de Associação e Classificação. Os algoritmos geram resultados de maneira diferente: a associação trabalha com *itemsets* e a formação de regras a partir destes; já a Classificação extrai modelos que descrevem classes importantes, tais modelos são chamados de classificadores e preveem rótulos de classes (HAN, 2012). Para este trabalho, utilizou-se algoritmos de Associação tais como Apriori e suas variações como Tid, Temporal, Híbrido, algoritmo SETM e AIS.

Os conhecimentos obtidos podem ser representados sob diversas formas como modelos lineares, árvores de decisões, regras de associação, regras de classificação e clusters. Com essas formas de representações, será adquirida uma visualização dos resultados do processo, os quais podem então ser averiguados e validados por um profissional da área. Portanto, a utilização de técnicas de mineração de dados e do processo de KDD se mostram bastante populares, sendo usados para a prevenção de doenças a partir dos dados de pacientes anteriores, encontrando padrões, além de classificação de doenças, monitoração e prognósticos (FARIAS et al, 2012).

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Este trabalho tem como objetivo aplicar algoritmos de Regras de Associação em uma base de dados na área da saúde, a qual contém ocorrências de prontuários médicos em Pato Branco, Paraná, para a obtenção de conhecimentos úteis para especialistas.

1.1.2 Objetivos Específicos

Os objetivos específicos são:

- Realizar levantamento bibliográfico de trabalhos similares para determinar e analisar algoritmos aplicáveis no mesmo domínio de problema;
- Analisar o banco de dados de prontuários médicos obtidos;
- Selecionar um conjunto de dados para geração de regras;
- Identificar, por meio da MD, regras de associação sobre problemas de saúde em prontuários médicos;
- Visualizar e analisar os resultados obtidos. A análise poderá ser feita com a apoio de especialistas ou comparação com outros trabalhos da literatura.

1.2 JUSTIFICATIVA

A aplicação de Mineração de Dados a uma base de dados na área da saúde traz informações antes obscuras dentro de tal base e que são de interesse para profissionais da saúde. Um exemplo pode ser encontrado no trabalho de Amin et al (2018) onde a MD auxiliou na busca por características relacionadas às doenças cardiovasculares.

Algumas modificações na base de dados são feitas, tais modificações serão a preparação para que a Mineração de Dados seja executada (WITTEN, 2011), tornando a base apropriada para encontrar resultados de interesse na resolução do

problema. As regras de associação resultantes de uma base de dados na área de saúde são analisadas para descobrir doenças, identificar prevenções ou indicar tratamentos de maneira eficaz.

Utilizando as informações encontradas nos dados de registros médicos, a realização da Mineração de Dados trará regras de associação que podem evidenciar novas informações como: as chances de ocorrência de uma determinada doença em um grupo de pessoas, ou enfermidades que ocorrem em um período do ano, tornando a prevenção de tais doenças efetivas.

Baseando-se na aplicação de regras de associação, um novo paciente que se encaixa nas condições de uma das regras será encaminhado para tratamento e diagnóstico com mais rapidez e eficácia.

A aplicação dos princípios do KDD e Mineração de Dados em conjunto de sistemas eletrônicos de atendimento ao paciente pode dar frutos a sistemas que ajudarão com mais facilidade a identificação de enfermidades e padrões que podem levar à descoberta das causas de determinadas doenças.

1.3 ESCOPO

O escopo do trabalho está no estudo do processo de KDD e aplicação de algoritmos de Regras de Associação na etapa de Mineração de Dados do KDD. Portanto, na base de dados que foi fornecida, já estará separado um conjunto onde os dados estarão selecionados e pré-processados. A aplicação do algoritmo e estudo das regras de associação serão o núcleo do trabalho. Tarefas como a filtragem dos dados, aquisição de novos dados ou desenvolvimento de softwares não fazem parte do escopo deste trabalho.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho foi estruturado para tratar, primeiramente, sobre seus dois conceitos principais: Prontuários Médicos e Descoberta de Conhecimento em Bases de Dados.

O capítulo 2 descreve os prontuários médicos e sua história, explorando também o conceito dos prontuários eletrônicos, seu uso no Brasil e a iniciativa e-SUS AB.

Durante o capítulo 3, são tratados os conceitos relacionados a Descoberta de Conhecimento em Bases de Dados, suas etapas, estratégias de pré-processamento e algoritmos de Mineração de Dados.

É também explorada a literatura, onde são tratados alguns trabalhos que abordam os temas relacionados a área da saúde por meio da Descoberta de Conhecimento, também é tratado o trabalho de Rodrigo Feuser, o qual este trabalho pretende dar continuidade.

São tratadas do capítulo 4 as ferramentas que foram utilizadas a realização do trabalho, com uma introdução aos programas necessários para a execução dos algoritmos e tratamento dos dados.

No capítulo 5 são discutidos os resultados da realização deste trabalho, por meio das regras de associação e também com o auxílio de gráficos e comparações.

2 PRONTUÁRIOS MÉDICOS

Neste Capítulo, pretende-se discutir prontuários médicos: o que são, seu histórico, como são gravados os dados, o prontuário eletrônico, e-SUS e seus sistemas, dando foco para o sistema WinSaude, da empresa IDS Desenvolvimento de Software e Assessoria LTDA.

O Capítulo é dividido em diversas seções as quais abordam os diversos tópicos relacionados a prontuários, descrevendo sua definição, origem e histórico traçando uma ligação com os prontuários eletrônicos. São listadas as vantagens e desvantagens do prontuário eletrônico, sua presença no Brasil. Também é realizado um estudo sobre o Sistema Único de Saúde (SUS), sua história, importância, os objetivos do SUS, bem como a descrição do prontuário eletrônico e-SUS. Finaliza-se o Capítulo com uma introdução ao software de prontuários eletrônicos IDS Saúde.

2.1 PRONTUÁRIO

O prontuário é um conjunto de documentos preenchido pelo profissional da saúde onde são registrados os cuidados prestados por profissionais ao paciente. São arquivados os prontuários após a consulta com o paciente, onde são encontradas informações como doenças, altura, exames realizados, condições de saúde até dados pessoais relacionados ao paciente, considerados informações de sigilo (GRISARD, 2000).

Segundo Marin (2003), a prática de relatos médicos feitos no papel datam do século V a.C, onde Hipócrates incentivou médicos a guardarem registros escritos, partindo de dois propósitos: replicar exatamente a maneira como a doença se desenvolveu no curso do tempo e o de examinar as possíveis causas de uma doença.

Durante o último quarto do século XVIII, na Europa, médicos se consolidaram como principais responsáveis das instalações hospitalares, antes dominada por instituições religiosas. Uma das principais mudanças que ocorreram nesse período foi um aumento na vigilância de pacientes. A vigilância era realizada com uso de etiquetas de identificação nos punhos dos pacientes, com fichas em cima de leitos, com o nome de doente e doença, registro de entradas e saídas de

pacientes, diagnósticos médicos, registro de enfermeiras, visitas, tratamentos e receitas fornecidas (SANTOS, 2007).

Conforme o tempo passou, cada vez mais foi difundida a ideia de manter os registros de pacientes e aos poucos, foi adotada a expressão “prontuário médico” para tais registros. Deste modo foi aprimorada a maneira a qual eles eram arquivados, a relação paciente-médico foi refletida com tais avanços.

Nos Estados Unidos, em 1880, foi formada a Clínica Mayo, por William Mayo e seus colegas de universidade. Observa-se que, nesse período, os empregados da clínica mantinham registros dos pacientes em forma cronológica de consultas, em um documento único, esse conjunto de anotações tornava a busca por informações específicas sobre um paciente no documento difícil. Durante 1907, William Mayo decide adotar um registro individual de consultas para cada paciente, assim se dá origem ao *prontuário médico individual centrado ao paciente* (MARIN, 2003).

Em 1920, houve uma movimentação nas clínicas Mayo para que alguns padrões fossem seguidos em prontuários. O objetivo era fazer com que os documentos seguissem determinadas diretrizes, e utilizar informações que os tornassem mais sistemáticos (MARIN, 2003). Esta abordagem demonstrou dificuldades, como o fato do dado clínico ser essencialmente heterogêneo. Por exemplo: dados de controle de sinais vitais teriam de ser atualizado periodicamente em planilhas, os registros e observações de um psicólogo se utilizariam de texto livre, ultrassonografia se apresenta na forma de imagens, entre muitas outras (MARIN, 2003).

Devido a situação de diversidade dos dados, foi introduzido o método de *prontuário orientado ao problema*. Idealizado por Lawrence Weed, em 1969, as informações dos pacientes eram registradas de acordo com uma estrutura chamada a partir de seu acrônimo inglês, SOAP (queixas; achados; testes e conclusões; plano de cuidado).

Um prontuário deverá ser, por obrigação, preenchido somente pelo médico e é responsabilidade do mesmo a manutenção dos registros. A única exceção seriam os hospitais de ensino, onde um aluno de medicina pode efetivamente cuidar de um prontuário, com a supervisão e sob responsabilidade de um médico. Apesar da guarda do prontuário ser responsabilidade das instituições hospitalares, o prontuário médico é propriedade do paciente e o mesmo tem direito à consulta e solicitação de cópia do documento (GRISARD, 2000).

Em 1960, foram iniciadas as primeiras experiências utilizando Sistemas de Informação, tendo como propósito a comunicação entre variados setores do hospital. Conforme se deu a utilização de sistemas, logo começaram a ser armazenadas informações relacionadas aos pacientes. Em 1972, um congresso foi feito a fim de estabelecer uma estrutura mínima para registros médicos. Logo, começaram a surgir os primeiros *Prontuários Eletrônico do Paciente* (PEP) (PATRICIO et al, 2011).

2.2 PRONTUÁRIO ELETRÔNICO DO PACIENTE

Durante os anos 90, o *Institute of Medicine* (IOM, 1997), conceitua o Prontuário Eletrônico do Paciente (PEP):

> "é um registro eletrônico que reside em um sistema especificamente projetado para apoiar os usuários fornecendo acesso à um completo conjunto de dados corretos, alertas, sistemas de apoio à decisão e outros recursos, como links para bases de conhecimento médico".

Além dessa definição, existem outras definições que possuem algumas diferenças do PEP, como o registro eletrônico do paciente, registro do paciente baseado em computador e registro eletrônico de saúde. Essas definições se relacionam com o fato de o prontuário consistir de digitalização de documentos, o que não pode ser considerado prontuário eletrônico, já que não engloba mudanças de comportamento e estruturação de informações (MARIN, 2003).

Segundo Marin (2003), o prontuário eletrônico é "um meio físico, um repositório onde todas as informações da saúde, clínicas e administrativas, ao longo da vida de um indivíduo estão armazenadas, e muitos benefícios podem ser obtidos deste formato de armazenamento". O armazenamento dos dados de prontuário eletrônico segue diversas diretrizes de padronização de dados e atualmente existem diversos tipos de padrões que direcionam os prontuários a possuírem uma linguagem comum mesmo com diferenças de *hardware* e *software* (COSTA, 2001).

De acordo com Costa (2001), são padrões como: identificação para pacientes e médicos (Cartão Nacional de Saúde, Número no Conselho Regional de Saúde), comunicação de mensagens entre sistemas (XML), conteúdo e estrutura dos registros clínicos (padrões fornecidos por órgãos como DATASUS),

representação dos dados clínicos (CID), confidencialidade e segurança, indicadores de qualidade e conjuntos de dados.

2.3 VANTAGENS E DESVANTAGENS DO PEP

De acordo por Sittig (1999), as vantagens do prontuário eletrônico são:

1. Acesso remoto: Profissionais podem, por meio da Web, acessar os dados de localidades remotas e vários profissionais podem ter o acesso simultaneamente;
2. Legibilidade: Registros escritos a mão podem tornar a leitura difícil, dados impressos não apresentam tais problemas;
3. Segurança: Apesar de preocupações com a perda de dados pelo mau funcionamento de sistemas, planos de prevenção contra erros e back-ups são uma opção muito mais segura contra a perda de dados do que registros escritos;
4. Confidencialidade: O acesso às informações contidas no prontuário pode ser restrito a usuários específicos. Registros de auditoria podem ser usados para a detecção de acessos não autorizados;
5. Flexibilidade: Há várias maneiras do usuário visualizar e interagir com os dados presentes, como ordem cronológica crescente ou decrescente, orientado a problema ou fonte;
6. Integração com outros sistemas: Em formato eletrônico, os dados podem ser enviados e acessados em computadores locais e também em áreas remotas de escolha do usuário;
7. Captura automática de dados: Podem ser utilizados equipamentos visuais localizados no hospital para que se evitem erros de digitação.
8. Processamento contínuo: Os dados podem ser verificados e atualizados automaticamente pelo software, procurando erros ou fazendo análise e interpretação de informações;
9. Assistência à pesquisa: Pesquisas podem ser feitas por meio de texto livre, palavras-chave afim de encontrar dados. Podem ser utilizadas pesquisas também parece verificar se dados foram salvos ou para levantamos estatísticos e estudos;

10. Diversas saídas de dados: Podem ser apresentados em diversas formas os dados, como impressos ou por e-mail. Também pode ser feito processamento em imagens;
11. Construção de vários tipos de relatórios: Alterações em fontes, cores e tamanhos na escrita podem ser feitas, para atentar aos pacientes a partes específicas do relatório. Também podem ser impressas imagens em conjunto com os dados;
12. Atualização constante: Sendo integrado, os dados do PEP estarão automaticamente disponíveis para todos os médicos de uma instituição.

Alguns obstáculos e desvantagens, relatados por Marin (2003), na implantação de um PEP que podem ser citados são:

1. Necessidade de investimento em hardware e software muito grande;
2. Usuários podem não se acostumar com os procedimentos necessários;
3. Possível ocorrência de sabotagem;
4. Sistema pode estar sujeito a falhas de hardware e software, tornando os dados inacessíveis pelo período em que as falhas estão sendo consertadas;
5. Dificuldade para a coleta de dados abrangente.

Outras dificuldades podem ser encontradas na utilização do PEP como seu impacto na relação médico-paciente, onde o contato direto entre ambos pode ser reduzido, também pode ser citado o aumento da carga de trabalho de um profissional da saúde, que terá de preencher uma grande quantidade de dados.

A utilização de PEPs diferentes pode acarretar em problemas também, com o possível impedimento no compartilhamento de informações, pela falta de padronização de sistemas (PATRICIO et al, 2011).

2.4 A UTILIZAÇÃO DO PRONTUÁRIO ELETRÔNICO NO BRASIL

No Brasil, o meio universitário começou a pensar em aplicações de um modelo de PEP nos anos 90. Dados os esforços, foram desenvolvidos modelos em diversas instituições em grandes centros urbanos do país. Pela necessidade de um padrão de informações e integração entre vários sistemas de informação na saúde.

Em 2002, foi proposto pelo Ministério da Saúde um mínimo de informações que devem estar em um prontuário médico (PATRICIO et al, 2011).

Definidos pela resolução da CFM nº 1.638/2002, os conteúdos que devem constituir um prontuário do paciente são:

- Ficha clínica com as seções: identificação, anamnese, exame físico, hipótese(s) diagnóstica(s) e plano terapêutico;
- Exames complementares: laboratoriais, exames anatomopatológicos, exame radiólogos, ultrassonográficos, etc.;
- Folha de evolução clínica;
- Folha de pedido de parecer;
- Folha de prescrição médica;
- Quatro TPR (temperatura - pulso – respiração);
- Resumo de alta/óbito.

Em julho de 2007, foi aprovado pelo Conselho Federal de Medicina (CFM) a resolução nº1.821/07, a qual continha normas relacionadas à utilização de sistemas de informação para a adquirir e manusear prontuários de pacientes, autorizando assim a eliminação do papel. A motivação foram os avanços da tecnologia, que apresentam novos métodos de armazenamento e transmissão de dados, considerando o grande volume de dados nos vários tipos de estabelecimentos de saúde (PATRICIO et al, 2011). De acordo com a resolução, ainda, para que a eliminação do papel seja possível, os prontuários devem atender os requisitos do “Nível de Garantia de Segurança 2”, permitindo a utilização de certificado digital padrão ICP-Brasil (Infraestrutura de Chaves Públicas Brasileira). Segundo o Instituto Nacional de Tecnologia da Informação (ITI, 2017), a ICP-Brasil é “uma cadeia hierárquica de confiança que viabiliza a emissão de certificados digitais para identificação virtual do cidadão” as entidades certificadas (pessoa, processo, servidor) é associado um par de chaves de criptográficas.

2.5 SUS, DATASUS E O PRONTUÁRIO ELETRÔNICO

Em 1988, através da promulgação da Constituição da República Federativa do Brasil, foi instituído o *Sistema Único de Saúde* (SUS), uma iniciativa pública para oferecer ao cidadão brasileiro acesso integral, universal e gratuito aos serviços de

saúde. Atualmente, o SUS realiza por ano cerca de 2,8 bilhões de atendimentos, desde procedimentos ambulatoriais simples até complexos, como transplante de órgãos. Com o SUS, a saúde passou a ser promovida e fazer parte dos planejamentos das políticas públicas (PENSESUS, 2018).

O Departamento de Informática do Sistema Único de Saúde (DATASUS) surgiu em 1991 com a criação da Fundação Nacional de Saúde (Funasa). Então, foi formalizada a criação e objetivos do DATASUS, que tem a responsabilidade de prover os órgãos do SUS sistemas de informações e auxílio de informática, nos projetos de planejamento, operação e controle. O DATASUS está presente em todas as regiões do país, por meio de Regionais que executam atividade de cooperação em informática nos principais estados do país (DATASUS, 2018).

Fazem partes das missões e objetivos do DATASUS (DATASUS, 2018):

- Fomentar, regulamentar e avaliar ações de informatização do SUS;
- Desenvolver, pesquisar e incorporar tecnologias de informática que tornem viável a implementação de sistemas necessários às ações da saúde;
- Definir padrões, diretrizes, normas e procedimentos para transferência de informações e contratação de serviços de informática em órgãos e entidades do Ministério;
- Definir padrões para a captação e transferência de informações em saúde, visando integração computacional entre bases de dados e dos sistemas desenvolvidos no âmbito do SUS;
- Manter o acervo de bases de dados necessárias aos sistemas de informações na saúde;
- Assegurar aos gestores do SUS o acesso a serviços de informática e bases de dados, mantidos pelo Ministério;
- Definir programas de cooperação técnica com entidades de pesquisa para transferência de tecnologia e metodologia em informação na saúde;
- Implementação do sistema nacional de informação em saúde.

De acordo com DATASUS (2018), o objetivo de criar um SUS que efetivamente atende a população exige organização e capacidades de gestão cada vez maiores. Para atingir esse objetivo, foi necessário integrar Sistemas de

Informação em Saúde (SIS) que contribuam com a integração entre diversos pontos e permitam interoperabilidade entre diferentes sistemas.

Baseado nos objetivos traçados pela criação do DATASUS, o Ministério da Saúde inicializou o projeto e-SUS, cujo nome faz referência ao SUS eletrônico. Seu objetivo é facilitar e contribuir com a organização do trabalho de profissionais da saúde, aspecto de grande importância para a qualidade de serviços prestados à população (DATASUS, 2018).

O e-SUS AB (Atenção Básica) é uma estratégia para reestruturar as informações de Atenção Básica em nível nacional, permitindo que a coleta de dados esteja inserida em atividades já desenvolvidas por profissionais. As possibilidades de utilização de tal estratégia estão adaptadas às realidades de cada município. Por exemplo, em Unidades Básicas de Saúde que contém computadores, é possível a utilização do sistema mesmo sem acesso à internet, fornecendo informações individuais (como idade, altura, dados pessoais, informações de morada, problemas de saúde, exames feitos) e funcionalidades que facilitam o dia a dia dos profissionais (PORTALDAB, 2018).

A informatização dos sistemas de saúde é prioridade na gestão do Ministério da Saúde, com objetivos que giram em torno de tornar o atendimento mais eficiente. O Prontuário Eletrônico do Cidadão (PEC) é ofertado gratuitamente pelo Ministério, reunindo histórico, dados, procedimentos realizados e resultados de exames realizado em pacientes do SUS, atendidos em Atenção Básica, também permitindo a consulta e disponibilidade de medicamentos, melhorando o atendimento ao cidadão (DATASUS, 2018). Podem ser visualizados na Figura 1 exemplos da realização de atendimentos por meio do PEC.

Figura 1 - PEC Ficha de Atendimento

Fonte: PORTALDAB (2018)

Conforme é possível observar, a janela de atendimentos contém dados relacionados aos cidadãos que irão comparecer, se o atendimento já foi realizado, data e hora de chegada, o profissional que atendeu e se o paciente compareceu ou não à consulta. Dos dados que são armazenados pelo PEC, pode-se observar dados gerais do cidadão, como nome completo, CPF, dados de moradia, informações de contato, agendamentos no e-sus. No sistema são registradas também informações de atendimentos (domésticos, individuais, odontológico), agendas e relatórios, cadastros domiciliares.

Por meio do Sistema e-sus AB, é liberado o compartilhamento de informações do PEC, de acordo com consentimentos do cidadão em relação à sua privacidade; para garanti-la, o sistema opta pelo modelo *opt-out* onde o cidadão pode optar em solicitar um bloqueio de compartilhamento de seus dados de atendimento (PORTALDAB, 2018).

O acesso dos dados disponíveis no sistema e-sus AB é feito unicamente a partir do login e senha, devendo ser responsabilidade do usuário o cuidado com a privacidade das informações. Pode ser feita a transmissão de dados do PEC a partir da instalação do software PEC Prontuário ou o PEC Centralizador, o qual faz papel intermediário na transmissão de informações (PORTALDAB, 2018).

2.6 SOFTWARE IDS E-SUS

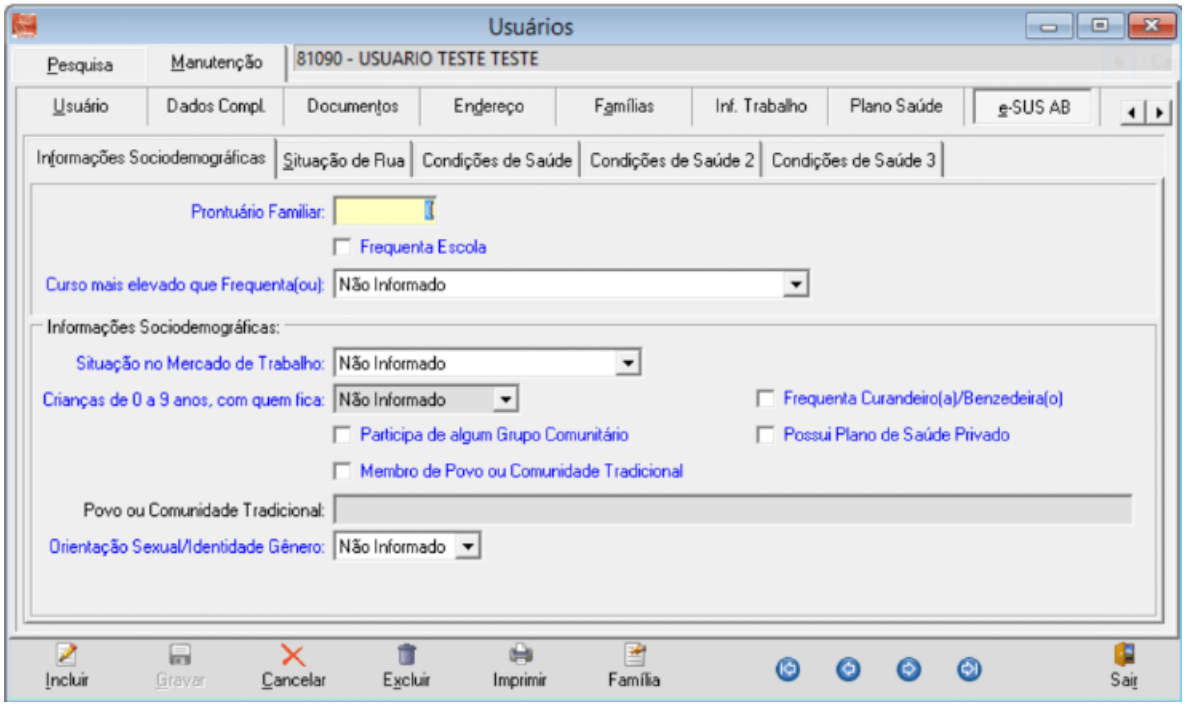
Fundada na cidade de Pato Branco, no polo tecnológico do Sudoeste do Paraná em 2003, a empresa IDS Desenvolvimento de Software e Assessoria LTDA atua no segmento de Gestão Pública Municipal. A empresa atua nas áreas de Gestão Municipal da Saúde, Assistência Social, Educação, Rural, auxiliando o desenvolvimento econômico e social de municípios (IDS, 2018).

Com sistemas voltados para a área de saúde, um dos projetos da empresa é o IDS Saúde. O sistema permite que informações sejam cadastradas e disponibilizadas para todos os setores da saúde, possibilitando a integração da Secretaria da Saúde. Utilizando a ferramenta sistematicamente, pode-se gerar o histórico individual do paciente. Também proporcionado para dispositivos móveis com GPS, para que acompanhamentos a domicílios sejam realizados. O IDS Saúde proporciona tomada de decisões rápida e precisa, através de sua interface operacional de sistema e pela interface de acesso do banco de dados. Além disso, os dados do IDS Saúde estão integrados e são transferidos automaticamente para os programas do DATASUS MS (IDS, 2018).

O Prontuário Eletrônico do Sistema WinSaude desenvolvido para o IDS Saúde possibilita a construção automática do histórico do paciente, a partir de atendimentos registrados, com a visualização de atendimentos anteriores (dados clínicos, medicamentos prescritos, encaminhamentos, diagnósticos, resultados de exames). Ele acessa automaticamente resultados de exames que foram realizados, permitindo visualizar informações de medicamentos disponíveis na farmácia e estoque na hora da prescrição médica, alertas para medicamentos que não fazem parte de farmácias municipais, entre outras informações.

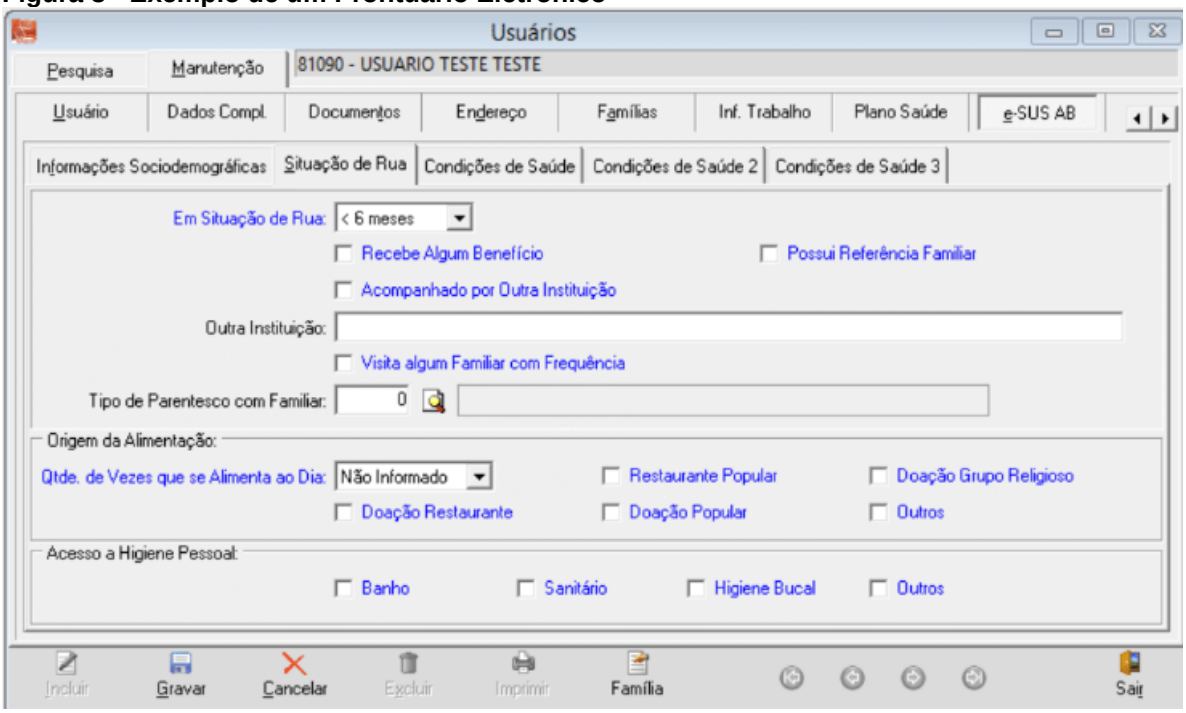
O Sistema Winsaude foi desenvolvido a partir de um modelo de entidades relacional e diagrama de caso de uso, com a utilização do IDE Delphi 7 e é integrado com vários módulos, usando um banco de dados baseado em SQL (Scructured Query Language, ou Linguagem de Consulta Estruturada), as tecnologias de banco de dados utilizadas pelo sistema são Firebird, Oracle, SQLServer e Postgres (NETTO; CAMARGO, 2013).

Figura 2 - Exemplo de um Prontuário Eletrônico



Fonte: IDS Winsaude (2017)

Figura 3 - Exemplo de um Prontuário Eletrônico



Fonte: IDS WinSaude (2017)

Na Figura 2 e na Figura 3 pode ser observada a utilização do Prontuário Eletrônico no Sistema WinSaude, com alguns exemplos de dados que são registrados no programa, como o registro de informações sociodemográficas e situação de moradia, fornecidas pelo paciente a fim de realizar o cadastro de usuário.

2.7 CONSIDERAÇÕES FINAIS

Este capítulo tratou-se de assuntos relacionados aos Prontuários Médicos, sejam em sua forma de papel, relatando as informações que são arquivadas dentro do mesmo, sua importância, história dentro do campo da medicina. Por fim, mostrou-se um estudo dos prontuários eletrônicos, com seu histórico, suas vantagens e desvantagens, informações gravadas no mesmo e sua utilização.

Observou-se também a utilização do Prontuário Eletrônico do Cidadão no Brasil, com as origens da utilização do prontuário eletrônico brasileiro, as iniciativas do e-sus AB e criação do DATASUS, norteando para o Sistema WinSaude, da empresa IDS Desenvolvimento de Software e Assessoria Ltda.

Com esse estudo, pode ser notada a importância de coletar e registrar informações relacionadas a atendimentos médicos, pacientes e doenças. Através da exploração das bases de dados que podem ser geradas com o registro de prontuários eletrônicos, muitas informações úteis para a área da saúde podem ser descobertas; partindo dessas finalidades, o capítulo 3 se dedicará ao estudo de algoritmos de Mineração de Dados (MD) usados neste trabalho.

3 MINERAÇÃO DE DADOS

Neste Capítulo será estudado o processo de Mineração de Dados, iniciando com a Descoberta de Conhecimento em Bases de Dados (KDD), sua definição, técnicas e etapas, tratando em detalhes os métodos utilizados em suas etapas. A Seção 3.1 conta com uma definição de dados, como funcionam, atributos. Também define o processo de KDD, esquematizando suas etapas e atividades. Na próxima Seção, 3.2, serão introduzidas as atividades do processo de KDD e a partir da Seção 3.3 serão discutidas, em detalhes, cada uma das etapas do processo.

Durante o processo de KDD, será discutido o tema principal desse capítulo: Mineração de Dados, onde será feita uma introdução ao processo de MD, bem como a descrição de diversos aspectos como a aprendizagem de máquinas, os algoritmos de associação, principalmente os que serão utilizados no trabalho: Apriori e Hotspot.

3.1 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

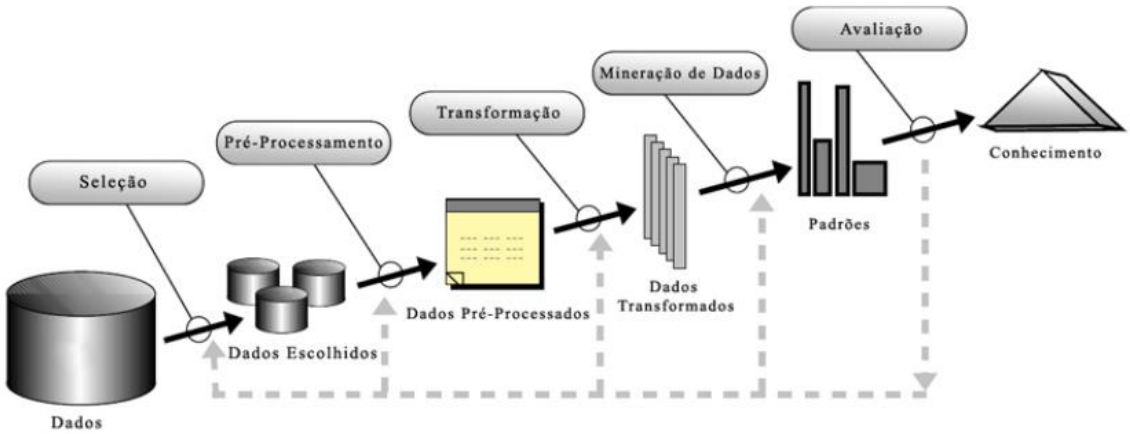
A Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases - KDD*) é como se chama a execução de processos que têm como objetivo a extração de informações antes ocultas e que são de interesse dentro de uma base de dados. As informações podem então se tornar conhecimento, por meio de avaliações e representações que indicarão se elas são de úteis ou não (FAYYAD et al, 1996).

Das preocupações do KDD, uma das principais é o desenvolvimento de métodos e ferramentas que ajudem a compreender dados. O problema básico abordado pelo KDD é o de mapear dados não tratados e representá-los de formas mais compactas ou úteis como, por exemplo, um relatório ou um modelo preditivo que estima valores que podem ser assumidos em casos futuros (FAYYAD et al, 1996).

O processo de KDD é dividido em várias etapas, que servem diversos propósitos na preparação de uma base para que seja obtido o conhecimento. Como ilustrado na Figura 4, o KDD se inicia com a base de dados sem alterações, passando por etapas como seleção, pré-processamento, transformação, execução

de um algoritmo de mineração de dados, análise e avaliação dos resultados, com a possibilidade de voltar à etapas anteriores.

Figura 4 - As etapas do KDD



Fonte: Fayyad et al (1996)

Começando com seus dados brutos, o primeiro passo do KDD é selecionar dados que serão de interesse para o projeto, eliminando informações irrelevantes para o estudo (FAYYAD et al, 1996). Para que tal passo ocorra, é preciso estudar a base de dados do problema, a fim de compreender o domínio da aplicação e identificar os objetivos de se executar o processo de KDD (FAYYAD et al, 1996).

A etapa de pré-processamento tratará de informações inconclusivas, erradas, duplicadas que estão entre os dados selecionados; após o pré-processamento tem-se a transformação dos dados, etapa onde os dados são preparados para a aplicação da MD (FAYYAD et al, 1996).

A Mineração de Dados é a etapa central do processo de KDD e principal foco deste trabalho. Nela é gerado o conhecimento a partir da procura de padrões dentro da base de dados: esse processo pode retornar à informação visualizada em gráficos, regras de associação ou árvores de decisão (HAN et al, 2012). Após a mineração, os dados são analisados e avaliados, geralmente por um profissional da área do estudo, caso determinados úteis, então obtém-se o conhecimento (FAYYAD et al, 1996).

Resultados do processo de KDD podem ser insatisfatórios, inconclusivos ou não oferecerem informações que são de fato úteis. Quando se vê problemas na avaliação dos resultados, podem ser repetidas etapas do processo, seja para selecionar os dados de uma maneira diferente, ou para executar outro algoritmo de Mineração de Dados. Diz-se que o processo de KDD é iterativo e interativo, devido à

possibilidade de ser repetido diversas vezes até que o resultado seja satisfatório e envolve diversas decisões realizada pelo usuário durante suas etapas (FAYYAD et al, 1996).

Segundo FAYYAD et al (1996) algumas aplicações de KDD em problemas reais notáveis estão: na ciência, com aplicações na astronomia, com destaque para o sistema SKICAT, que realiza análise de imagens, classificação e catálogo de astros; também pode ser visto em diversos ramos de negócios, como marketing, finanças (investimento), detecção de fraude, telecomunicações. Em outras áreas, um exemplo da utilização pode se dar com o sistema da IBM *ADVANCED SCOUT*, que utiliza técnicas de mineração de dados para a ajudar técnicos da NBA na análise de resultados e dados de jogos de basquete (FAYYAD et al, 1996).

3.2 SELEÇÃO DE DADOS

A fase de Seleção é considerada a etapa onde é criada a familiaridade com o domínio onde a aplicação será executada, sendo a primeira no processo de descoberta de conhecimento. Um conjunto de dados é definido a partir da seleção de variáveis e registros que serão escolhidos para as próximas fases do processo (PRASS, 2012). Fazem parte do processo: descobrir quais dados estão disponíveis, obter dados adicionais para descoberta e integrar todos os dados que serão utilizados na descoberta de conhecimento em um conjunto de dados, incluindo os atributos que serão considerados no processo (ROKACH; MAIMON, 2010).

Os responsáveis pelos estudos dentro do domínio devem ser capazes de entender e adquirir uma finalidade com as informações que estarão dentro do conjunto de dados. Para isso, deve ser criada uma familiaridade com as informações que possam ser úteis para o mesmo na hora de lidar com os dados. Deve-se traçar o objetivo para o qual o processo de Descoberta de Conhecimento será utilizado e, a partir desse objetivo e de informações cedidas pelos responsáveis do estudo, separar as instâncias e atributos que virão a interessar no projeto (HAN et al, 2012).

É necessário reunir os dados em um conjunto de instâncias, em um caso real isso pode significar realizar um processo onde dados de diferentes departamentos são integrados ao conjunto onde será realizado o estudo. Os dados de fontes diferentes podem vir com atributos alterados, valores registrados de

maneira diferente; uma das complicações da integração entre os dados são os erros que surgir por causa do processo, além de reunidos, integrados, os dados também devem passar pela etapa de limpeza de dados (WITTEN, 2011).

Segundo Rokach e Maimon (2010), a importância dessa etapa está no fato de que a Mineração de Dados descobre informações a partir dos dados disponíveis. Essa é a base para a construção da representação de conhecimento, se atributos importantes não estão presentes, o estudo pode falhar: por essa perspectiva, quanto mais atributos, melhor. Porém, coletar, organizar e operar em cima de repositórios complexos de dados pode ser caro e apresenta uma oportunidade recompensadora de entender acontecimentos. Essa recompensa é uma demonstração dos aspectos interativos e iterativos do KDD.

3.3 LIMPEZA DOS DADOS

Mineração de Dados geralmente é aplicada em dados que foram coletados por motivos diferentes do estudo que estará sendo aplicado. Na maioria dos casos, a resolução de problemas em qualidade de dados não é uma opção. Detectar e resolver problemas de Qualidade de Dados ou o uso de algoritmos tolerantes à qualidade pobre são os focos desse processo na Mineração de Dados. O processo de detectar e resolver problemas relacionados à qualidade dos dados é chamado de Limpeza de Dados (TAN,2009).

Um problema comum é que o valor registrado difere do valor verdadeiro, de certa forma. Para atributos contínuos, a diferença numérica entre um valor medido e seu valor verdadeiro é chamada de erro. Coleta de erros em dados se refere à problemas como a omissão de dados ou atributos e objetos incluídos erroneamente (TAN,2009).

É comum que um objeto possua um ou mais valores de atributos faltantes. Em alguns casos, a informação não foi coletada (pessoas podem não informar idade, por exemplo), em outros, a informação não pode ser aplicada a todos os objetos (TAN,2009).

Dados podem conter valores inconsistentes, como endereços digitados erroneamente, seja por digitação ou por uma leitura errada do dado escrito em papel. Alguns tipos de inconsistências são fáceis de identificar, porém em alguns

casos será necessário consultar fontes externas para tal identificação (TAN, 2009). Já identificada a inconsistência, é possível corrigi-la, caso de simples identificação, como um dado de altura negativo. Em dados como códigos de produtos, onde a checagem necessitaria de uma lista de códigos válidos, mostra-se importante a consulta de fontes externas e informações adicionais (TAN, 2009).

3.4 PRÉ-PROCESSAMENTO

O pré-processamento pode ser definido como a etapa do KDD onde podem ser feitas aplicações que tornam os dados adequados para a etapa de transformação em relação a tempo, custo e qualidade (TAN, 2009). Abaixo estão algumas dessas abordagens:

- Agregação;
- Amostragem;
- Redução de dimensionalidade;
- Seleção de subconjuntos de características;
- Criação de características;
- Discretização e binarização;
- Transformação de atributos.

As técnicas, descritas em maiores detalhes nas próximas seções, podem ser descritas em duas categorias: objetos que serão selecionados para análise e criação/alteração de atributos (TAN,2009).

3.4.1 Agregação

Agregação é a combinação de dois ou mais objetos em um só, pode também ser visto como a eliminação de atributos, ou redução do número de valores para um atributo só (TAN,2009).

Tabela 1 - Registro de atendimentos médicos

Atendimento	Data de Consulta
0991	04/06/2018
0992	08/07/2018

Fonte: Autoria própria

Tabela 2 - Registro de atendimentos médicos com dados agregados

Atendimento	Data de Consulta
0991	06
0992	07

Fonte: Autoria própria

Por exemplo: a Tabela 1 representa um registro de atendimentos médicos dentro de uma unidade de saúde, onde suas datas são registradas através dos 365 dias do ano, esse atributo, Data de Consulta, poderia ser reduzido para 12 meses.

3.4.2 Amostragem

Amostragem é usada para a seleção de um subconjunto de dados a ser analisado. Esta técnica é usada há tempo na Estatística em investigações preliminares de dados e análises finais. Na Mineração de Dados ela é usada a devido à dificuldade e custo de se processar todos os dados. Utilizar um algoritmo de amostragem pode reduzir o tamanho dos dados até um ponto onde um algoritmo melhor e mais caro pode ser usado (TAN,2009).

É necessário escolher o tamanho da amostra após a escolha de uma das técnicas de amostragem. Amostras de tamanho grande possuem uma chance alta de trazerem dados representativos, porém eliminam boa parte dos motivos de se fazer amostragem. Já com amostras menores, padrões podem ser perdidos ou até estarem representados de maneira errônea, geralmente se opta por um tamanho médio de amostragem, onde as características ainda são mantidas (TAN, 2009).

3.4.3 Redução de Dimensionalidade

Dados podem possuir grandes quantias de atributos, essa quantia de atributos é chamada de dimensionalidade. De acordo com Tan (2009), o termo Redução de Dimensionalidade é reservado para técnicas que reduzem a dimensionalidade por meio da criação de um novo atributo a partir da combinação de atributos antigos. Algoritmos de Mineração de Dados podem se beneficiar da Redução de Dimensionalidade, que pode eliminar características irrelevantes a análise e reduzir ruídos. Outro benefício da redução é a criação de um modelo compreensível por envolver uma quantia menor de atributos; também pode facilitar a visualização dos dados. Com uma redução em dimensionalidade, a quantia de tempo e memória para executar a mineração pode diminuir consideravelmente.

3.4.4 Seleção de Subconjuntos de Características

Outra maneira de reduzir a dimensionalidade é selecionar apenas um subconjunto das características, tal processo pode resultar em perda de informações, porém não será o caso se características redundantes ou irrelevantes estão presentes (TAN, 2009). Por exemplo, um caso onde pretende-se analisar uma base de dados onde são registrados desempenhos de estudantes em uma escola, sendo a informação desejada a média das notas entre os alunos, seus nomes ou ID seriam irrelevantes para a análise (TAN, 2009).

Características redundantes podem duplicar muita ou toda a informação contida em um ou mais atributos (TAN, 2009). Tais características podem influenciar a exatidão das saídas resultantes do algoritmo, visto que os dados duplicados ainda representam informações para o algoritmo, alterando o resultado final do mesmo.

3.4.5 Criação de Características

A partir dos atributos originais, pode-se criar um novo conjunto de atributos, capturando as informações importantes de um conjunto de dados em uma maneira muito mais eficaz. Também pode reduzir o número de atributos, trazendo as mesmas vantagens das abordagens relacionadas a dimensionalidade (TAN,2009).

3.4.6 Discretização e Binarização

Algoritmos de Mineração de Dados, principalmente alguns algoritmos de Classificação, precisam de dados em forma de atributos categóricos. Algoritmos que retornam padrões de Associação requerem atributos binários. A discretização é o processo de se transformar um atributo contínuo em um atributo categórico. Já a binarização pode ser descrita como o processo de transformar atributos contínuos ou discretos em atributos binários (TAN, 2009).

3.4.7 Transformação de Atributos

Uma Transformação de Atributos é uma transformação aplicada em todos os valores de um atributo. Para cada objeto, a transformação é aplicada para o valor do atributo daquele objeto (TAN, 2009).

Dos métodos de transformação, um dos mais comuns é a Padronização ou Normalização, sendo que ambos os nomes são utilizados dentro da área de MD para o mesmo processo. Segundo Tan (2009), a meta da Padronização/Normalização é fazer um conjunto de valores possuir uma propriedade em particular. Um exemplo é a “Padronização de uma variável” em estatística: se x representa o valor de um atributo, m a média dos valores contidos em um atributo, e s o desvio padrão, então a transformação $x' = (x - m)/s$ criaria uma variável nova que possui média 0 e desvio padrão 1 (TAN, 2009)

3.5 TRANSFORMAÇÃO DOS DADOS

Com os dados selecionados e pré-processados, eles necessitam ser armazenados e formatados de forma onde algoritmos de aprendizado de máquina, usados na etapa de Mineração de Dados, possam ser aplicados. É comum encontrar dados dispersos em vários sistemas operacionais ou bancos de dados, os mesmos devem ser agrupados em um repositório únicos (ROKACH; MAIMON, 2010). Em alguns casos, a transformação de dados é aplicada antes da etapa de Seleção dos Dados (HAN et al, 2012).

Os dados devem ser armazenados em um arquivo onde estão representados todos os atributos de entrada, são formatados em um arquivo ARFF (*Attribute Relation File Format*). O arquivo ARFF não especifica os atributos que serão utilizados, apenas os valores que cada um toma e suas instâncias, isso significa que o mesmo arquivo pode ser utilizado para estudar como cada um dos atributos será previsto. Também já pode ser usado para gerar regras de associação e clusterização. No arquivo *arff*, cada instância ocupa uma linha e cada valor de atributo é separado por uma vírgula, no caso de valores faltantes, é usado um ponto de interrogação valor do atributo (HAN el at, 2012).

3.6 MINERAÇÃO DE DADOS

Vista como o núcleo do processo de KDD, a Mineração de Dados consiste no processo de descoberta de padrões relevantes e da descoberta de conhecimento dentro de uma grande quantia de dados (HAN et al,2012).

A Mineração de Dados visa encontrar modelos que encaixam nos padrões que foram encontrados dos dados observados. Os modelos tomam o papel de conhecimento inferido, onde se esse conhecimento é de interesse, válido ou não será determinado pelo processo geral da fase interativa do KDD, onde se requer julgamento humano (FAYYAD et al, 1996).

Geralmente, somente uma parcela das informações extraídas pela etapa de Mineração de Dados pode ser considerada interessante, embora o processo seja capaz de gerar vários padrões ou regras. O algoritmo escolhido para o processo de Mineração de Dados definirá a maneira com que a saída será representada, junto com a influência das etapas anteriormente estudadas, diversas maneiras de se dispor os padrões encontrados podem ser geradas. Portanto qualquer alteração no processo pode trazer uma saída diferente, fazendo com que a avaliação dos resultados seja importante para determinar quais dados são realmente interessantes (HAN et al, 2012).

O que faz um padrão se tornar interessante são, segundo HAN et al (2012), ser de fácil compreensão para pessoas, válido em novos dados com algum grau de certeza, potencialmente útil ou novo; um padrão também pode ser interessante se

confirma uma hipótese do usuário. O padrão interessante pode ser considerado conhecimento (HAN et al, 2012).

Segundo Tan (2009), tarefas da Mineração de Dados podem ser divididas em dois tipos: descritiva e preditiva. O objetivo das tarefas preditivas é o de prever valores relacionados a um atributo em particular baseado em valores de atributos anteriores. São de natureza preditiva Classificação, Regressão, Detecção de Desvios. O atributo a ser previsto é conhecido como alvo ou variável dependente, enquanto o atributo utilizado para a previsão é chamado variável independente ou explanatório (TAN,2009). As tarefas descritivas têm como objetivo derivar padrões que consigam resumir as relações internas entre dados. São exploratórias por natureza e requerem técnicas de pós-processamento para validar seus resultados. Exemplos de tarefas descritivas são Agrupamento e Regras de Associação (TAN,2009).

3.6.1 Aprendizagem de Máquina

Sendo um domínio com diversos tipos de aplicações, a Mineração de Dados incorporou várias técnicas interdisciplinares em seus processos, essa natureza interdisciplinar contribui para o sucesso de pesquisas e desenvolvimento em MD (HAN et al, 2012).

A Aprendizagem de Máquina se refere a área da Inteligência Artificial que investiga as maneiras em que um computador aprende (ou melhora sua performance) a partir de dados. Uma grande área de pesquisa é como computadores podem aprender a reconhecer padrões complexos e fazer decisões baseadas em dados automaticamente (HAN et al 2012). Apesar de parecidos conceitos, a Aprendizagem de Máquina não representa o processo de MD, ela oferece métodos que são utilizados durante a etapa de MD no KDD (FAYYAD et al, 1996). Dentre outras formas, a aprendizagem pode ser supervisionada ou não-supervisionada.

A Aprendizagem Supervisionada, segundo Han et al (2012), pode ser considerada um sinônimo para Classificação. A supervisão vem de exemplos marcados no conjunto de dados, que serão utilizados para treinamento do algoritmo.

São exemplos de Aprendizagem Supervisionada árvores de decisão, redes neurais, classificação de bayes, algoritmos genéticos.

A aprendizagem não-supervisionada, segundo Han et al (2012), pode ser definida como sinônimo para Clusterização. O processo de aprendizado não é supervisionado pois os exemplos de entrada não estão rotulados com uma classe. Tipicamente, Aprendizagem Não-Supervisionada é utilizada para definir classes dentro de um conjunto de dados. Como os dados de treino não são rotulados, o modelo resultante não terá definições semânticas das classes encontradas. Exemplos de Aprendizagem Não-Supervisionada são K-means e Clusterização.

3.6.2 Associação

A metodologia conhecida como análise de associação é útil para a descoberta de relações interessantes entre registros em uma grande quantia de dados. Estas relações podem ser representadas através de Regras de Associação (SE-ENTÃO) ou conjuntos de itens frequentes (TAN, 2009).

Por exemplo, é possível extrair a seguinte regra a partir de uma base de registros de compras de supermercado: Fraldas ⇒ Cerveja (SE há uma compra de fraldas, ENTÃO há uma compra de cerveja.)

Isso significa que existe um possível relacionamento entre a venda de fraldas e cerveja porque muitos clientes compram a cerveja junto com fraldas. A partir desse tipo de regras vendedores podem identificar oportunidades para atrair consumidores (TAN, 2009).

Um conjunto de itens se refere a coleção de zero ou mais itens em um conjunto. Se um conjunto de itens possui *k* elementos, ele é um conjunto de *k*-itens (TAN, 2009). Por exemplo, o conjunto *{Roupão, Cobertas, Suco}* seria um conjunto de 3-itens

Uma regra de associação é uma expressão de inferência no formato $X \Rightarrow Y$ onde X e Y são conjuntos de itens disjuntos. A partir de confiança e suporte, pode-se determinar se uma regra é forte ou não. Confiança determina o quanto itens em Y aparecem em instâncias que contém X e suporte determina o quão aplicável é uma regra em cima de um conjunto de dados (TAN, 2009).

$$Suporte, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \quad (1)$$

$$Confiança, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

Em algumas situações regras com um valor de confiança alto podem não ser totalmente confiáveis, pois a medida de confiança não leva em consideração o suporte contido na parcela consequente da regra. Uma maneira de remediar este problema é a utilização de uma métrica chamada *Lift* (TAN, 2009).

$$Lift = \frac{c(A \rightarrow B)}{s(B)} \quad (3)$$

A métrica Lift se utiliza das duas medidas, suporte e confiança, e seu resultado é a razão entre as duas medidas. Com isso, não só a confiança de uma regra é levada em conta, como também sua posição para com o conjunto de dados completo (TAN, 2009).

Mineração de Regras de Associação pode ser vista como a execução de duas atividades, segundo Han et al (2012):

1. Encontrar todos os conjuntos de itens frequentes;
2. Gerar regras de associação fortes a partir dos conjuntos de itens, que satisfarão suporte e confiança mínimos.

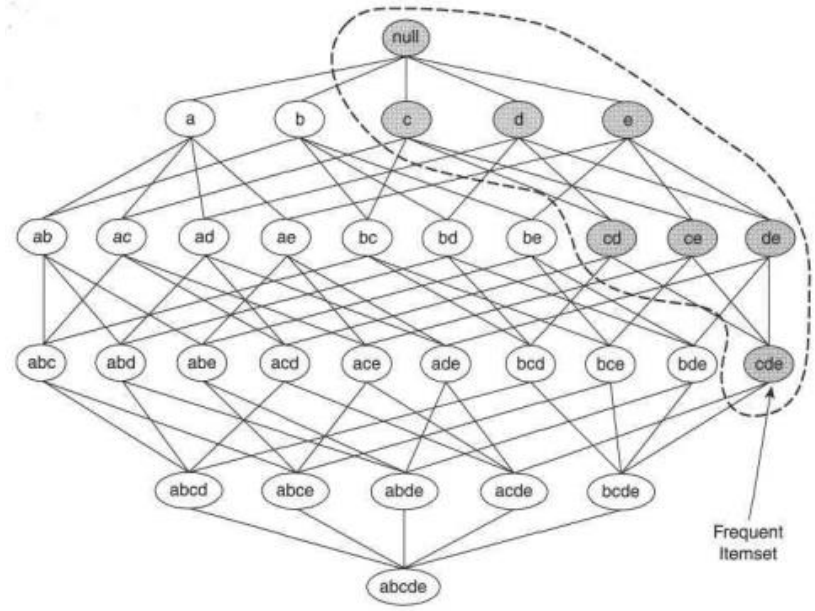
Como o número de conjuntos de itens frequentes pode aumentar exponencialmente, existem várias maneiras de se reduzir o custo computacional da geração de conjuntos de itens frequentes (TAN, 2009).

3.6.3 Teorema de Apriori

O teorema de *Apriori*, visando diminuir o número de conjuntos de itens candidatos, propõe que se um conjunto é frequente, então todos os seus subconjuntos também são (HAN et al, 2012).

A Figura 5 ilustra o teorema de Apriori, onde o conjunto {*c,d,e*} é um conjunto frequente de itens. Como todo conjunto que contenha {*c,d,e*} é um conjunto frequente de itens, ele também deve conter seus subconjuntos {*c,d*}, {*d,e*} ou {*c*}. Portanto, se {*c,d,e*} é frequente, então todos os subconjuntos de {*c,d,e*} serão frequentes também.

Figura 5 - O teorema de Apriori ilustrado



Fonte: TAN (2009)

Porém, se um subconjunto como {a,b,c} não for frequente significa que todos os seus subconjuntos não são frequentes também, logo, todo o subgrafo {a,b,c} será podado do grafo (TAN, 2009).

O algoritmo de *Apriori* foi o primeiro algoritmo de mineração de dados a utilizar o método de poda mostrado acima, conhecido como poda baseada em suporte, para controlar o crescimento exponencial dos crescimentos dos conjuntos de itens (TAN, 2009).

A estratégia do algoritmo Apriori, então, pode ser descrita como a identificação de conjuntos de itens frequentes, após a formação de conjunto, serão construídas regras de associação (MARIANO, 2011). A utilização da poda baseada em suportes vem da escolha de conjuntos frequentes candidatos, onde um conjunto não frequente terá seus subconjuntos descartados da análise (TAN, 2009).

A seguir, no Código 1, está contido o pseudocódigo da função principal do algoritmo Apriori.

Código 1 - Algoritmo Apriori

```
1.  Entrada: Uma base de dados D e o valor de suporte mínimo min_sup.
2.  Saída: O conjunto L com todos os itemsets frequentes.
3.  Função apriori-main(D, min_sup)
4.  L1 = {conjunto dos itemsets frequentes de tamanho 1 contidos em D};
5.  para (k = 2; Lk−1 ≠ ∅; k++)
6.      Ck = apriori-gen(Lk−1);
7.      para todas transações t ∈ D fazer
8.          Ct = subset(Ck, t);
9.          para todos candidatos c ∈ Ct fazer
10.             c.count++;
11.         fim para
12.     fim para
```

Fonte: MARIANO (2011)

O algoritmo faz uma varredura pela base de dados *D*, a fim de gerar um conjunto de itens frequentes tamanho *L* para determinar o suporte de cada item com base em uma medida de suporte mínimo (*min_sup*). No fim dessa varredura, será gerado o conjunto de itens frequentes tamanho 1 contidos do conjunto de dados (TAN, 2009);

Será tratada a geração de conjuntos candidatos de tamanho *k* a partir de conjuntos frequentes de tamanho *k-1* com a função apriori-gen, gerando o conjunto *C*, o que poderá ser observado no Código 2 (TAN, 2009).

Código 2 - Geração de conjuntos candidatos

```
1.  Função apriori-gen(Lk)
//  Passo 1 (join)
2.  para cada itemset l1 ∈ Lk−1 fazer
3.      para cada itemset l2 ∈ Lk−1 fazer
4.          se (l1[1] = l2[1] ∧ l1[2] = l2[2] ∧ ... ∧ l1[k-1] < l2[k-1]) então
5.              c = l1[1] . l1[2] ... l1[k-2] . l1[k-1] . l2[k-1];
6.              adicione c em Ck;
7.          fim se
8.      fim para
9.  fim para
//  Passo 2 (prune)
10. para todos candidatos c ∈ Ck fazer
11.     para todos (k-1)-subsets s ⊂ c fazer
12.         se (s ∉ Lk−1) então
13.             delete c de Ck;
14.         fim se
15.     fim para
16. fim para
17. retorne Ck;
```

Fonte: MARIANO (2011)

O cálculo de suporte dos candidatos será realizado através da função *apriori-gen* (Código 2), que recebe como entrada um conjunto composto por todos os grupos de itens frequentes (*itemset*) e retorna o conjunto de candidatos *C*. Essa

função é dividida em duas: *join* e *prune*. O primeiro passo representa a junção dos conjuntos presentes em i em possíveis candidatos. Já o segundo passo representa a poda, onde são removidos todos os conjuntos não frequentes (MARIANO, 2011).

São determinados todos os conjuntos candidatos *C* que estão contidos em cada transação *t* (TAN, 2009).

Código 3 - Geração do conjunto de regras

```
1.  Entrada: Um conjunto de itemsets L e a confiança mínima da regra min_conf.
2.  Saída: O conjunto de regras R.
3.  Função ap-genrules(L, min_conf)
4.  para todos k-itemsets ∈ L fazer
5.      para (i = k-1; i ≥ 1; i−−)
6.          para todos i-itemsets ⊂ k-itemset fazer
7.              conf = suporte(k-itemset) / suporte(i-itemset);
8.              se (conf ≥ min_conf) então
9.                  adicione i-itemset → (k-itemset − i-itemset) em R;
10.             fim se
11.         fim para
12.     fim para
13. fim para
14. retorne R;
```

Fonte: MARIANO (2011)

Gerado o conjunto *L* com todos os conjuntos de itens frequentes, é dado início a fase de extração de regras de associação (MARIANO, 2011). Essa etapa está exemplificada no Código 3, onde o conjunto *L* é recebido junto a confiança mínima (*min_conf*) a qual foi requisitada para a formação das regras. Para todos os conjuntos de itens (*k-itemsets*) presentes em *L*, são extraídos todos os subconjuntos de itens (*i-itemsets*). Para cada subconjunto, é calculada a confiança da regra que será gerada. Caso satisfaça a confiança determinada anteriormente, a regra é então adicionada ao conjunto de saída *R* do algoritmo (MARIANO, 2011).

3.6.4 Algoritmo Hotspot

A análise Hotspot escava dados sistematicamente e, procurando pelos segmentos de maior interesse, detecta relacionamentos, dependências, associações, variáveis e interações entre dados para prover informações estatísticas para segmentação de perfis em um conjunto de registros (ROSELLA, 2005).

O processo de análise do Hotspot consiste em (ROSELLA, 2005):

- Segmentação: dividir os dados em segmentos;

- A partir dos segmentos, gerar perfis;
- Realizar os intervalos numéricos em formato *drill-down;*
- Seleção de variáveis que serão usadas para segmentação de perfis;
- Ordenar segmentos baseados em uma pontuação;
- Visualização dos resultados.

O algoritmo de regras de associação baseado em segmentação de perfis Hotspot foi desenvolvido por Mark Hall em 2010 inspirado pelos princípios de análise Hotspot. Foi implementado na linguagem de programação Java e está disponível dentre os algoritmos de associação da ferramenta WEKA. A aplicação do algoritmo em uma base de dados gera regras de associação, que podem ser visualizadas em uma estrutura parecida com a de uma árvore de decisão (HALL, 2010).

Utilizado para extrair informações de um alvo específico dentro da base de dados, as regras do algoritmo são geradas com base em um atributo e um valor escolhido para a resolução do problema, afim de maximizar ou minimizar o valor de interesse do atributo. São geradas regras de associação a partir do atributo escolhido (HALL, 2010).

O algoritmo HotSpot recebe como entrada uma base de dados, o suporte mínimo e o valor alvo pelo qual serão geradas as regras. A partir desses valores, o algoritmo realizará uma busca gulosa pela base de dados, a fim de maximizar/minimizar o valor de interesse do atributo e valor selecionados para a execução do algoritmo, encontrando as regras onde o valor desejado está presente (HALL, 2010).

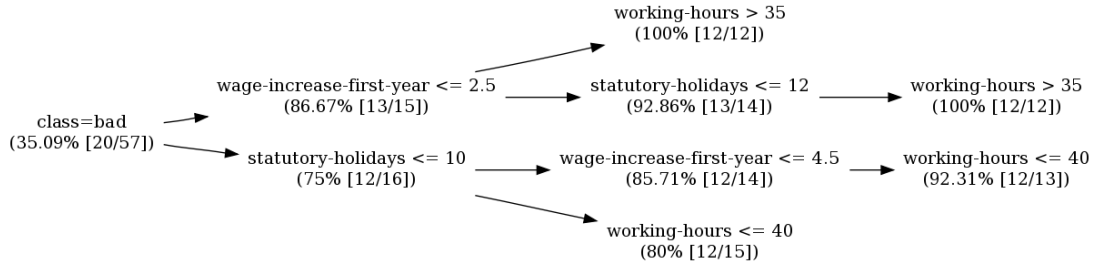
A base de dados é percorrida e é gerado um grupo de regras baseando seu lado direito/consequente no valor alvo determinado pela entrada (HALL, 2010). A partir desse conjunto de regras, são realizados testes nas regras a fim de encontrar as regras que possuem maior concentração do valor alvo (i.e. regras onde a porcentagem de confiança do valor acontecer é maior).

Depois de avaliadas pela porcentagem e devidamente ordenadas por valor de confiança, são eliminadas do conjunto de saída as regras que não satisfazem o valor de suporte mínimo determinado pela entrada. Por fim, caso seja desejado pela execução do arquivo, as regras finais são estruturadas na forma de uma árvore.

O exemplo com saída representada na Figura 6 utiliza uma base de dados de negociações trabalhistas do Canadá, seus atributos são o aumento de salário

depois do primeiro ano (*wage increase first year*), contagem de feriados nacionais (*statutory holidays*), horas de trabalho semanais (*working hours*) e contratos (*class*). A saída pode ser visualizada na forma no grafo de árvore de onde podem ser extraídas as regras de associação (PENTAHO, 2010).

Figura 6 - Saída do algoritmo Hotspot



Fonte: PENTAHO (2010)

A execução do algoritmo focou seus resultados nas regras que possuíam como consequente o atributo *class*, que se refere ao resultado do contrato e seu valor '*bad'* que indica os contratos que foram rejeitados. Os valores indicam a porcentagem de confiança de cada elemento e os números as instâncias na base de dados nas quais aparecem.

Uma das regras de associação que podem ser extraídas da árvore, por exemplo, (*statutory holidays* <= 10, *working hours* <=40 -> *class* = *bad*), indica que contratos onde trabalham-se 40 horas ou menos e possuem 10 ou menos feriados nacionais são recusados.

3.7 COMPARATIVO ENTRE APRIORI E HOTSPOT

As diferenças de desempenho entre o Apriori e o Hotspot estão principalmente nas situações onde são utilizados. Onde o Apriori gera suas regras a partir de todos os conjuntos de itens, o algoritmo Hotspot tem sua aplicação focada em apenas um valor, traçando suas regras de associação

Para pesquisas e experimentos dentro da base de dados completa, a utilização do Apriori se mostra mais eficiente, enquanto pesquisas que buscam traçar perfis ou segmentar valores dos registros encontrados na base de dados irão se beneficiar do uso do Hotspot, seja para encontrar os valores máximos ou mínimos de um atributo (HALL, 2010).

Comparado com o Hotspot, a maior vantagem do Apriori é a geração de regras com diversos valores no lado direito de suas regras, o que resulta em uma exploração mais ampla dos conjuntos de itens presentes nas bases de dados e um número maior de regras para serem analisadas. Os resultados do algoritmo Apriori são mais satisfatórios para aplicações onde o conhecimento de interesse está relacionado com todas as classes presentes nos registros de uma base de dados.

3.8 AVALIAÇÃO DOS RESULTADOS

São avaliados e interpretados os resultados da etapa de Mineração de Dados, sejam eles regras, árvores de decisão, estatísticas, baseando-se no problema o qual foi definido durante as primeiras etapas do processo de KDD. O foco desta etapa está em verificar se o modelo desenvolvido é compreensível, útil e coincide com as metas e objetivos traçados pelo projeto (ROKACH; MAIMON, 2010).

Caso não haja um resultado satisfatório, o projeto pode retornar à outras etapas do KDD presentes na Figura 4 - As etapas do KDD. Segundo Rokach e Maimon (2010), são ações comuns em uma nova iteração do KDD modificar o conjunto inicial de dados e/ou alterar o algoritmo de MD que será aplicado. Com o sucesso desta etapa, o conhecimento obtido pode ser documentado para uso em aplicações futuras (ROKACH; MAIMON, 2010).

3.9 TRABALHOS RELACIONADOS

A partir de pesquisas realizadas nas bases de dados Springerlink, Ieeexplorer e Sciencedirect, foram selecionados alguns artigos e trabalho relacionados aos temas tratados nesse trabalho a fim de enriquecer o embasamento teórico do mesmo, provendo uma base de trabalhos anteriores presentes na literatura.

Gomez-Verján, Gutiérrez-Robledo (2018), estudam o desafio entre a aplicação de Mineração de Dados e problemas relacionados a doenças crônicas providas pelo envelhecimento. Utilizando de ferramentas digitalizadas por pesquisas clínicas, a aplicação de MD é feita com o fim de descobrir informações que ajudem a abordar as complexidades do envelhecimento.

Amin et al (2018), conduz experimentos com a aplicação de Mineração de Dados para previsão de doenças cardiovasculares, a fim de encontrar as características frequentes na previsão de problemas cardíacos e melhorar a exatidão dessas previsões, desenvolveu modelos preditivos, com técnicas de classificação como: k-nn, árvores de decisão, redes neurais. Resultados mostraram os melhores algoritmos com uma exatidão de 84,7 % em previsões.

Moreira e Namen (2018), aplicando a MD dentro de uma base de dados de diagnósticos de pacientes com problemas cognitivos, encontrou modelos preditivos através de algoritmos de classificação para Alzheimer e deficiências cognitivas, demonstrando que tais problemas são causados principalmente em idosos com pouca escolaridade, onde diagnósticos prematuros são um dos desafios da medicina.

Hannun e Andrade (2018) fazem um estudo de técnicas de Aprendizagem de Máquinas que foram utilizadas no campo do transplante renal, que já possui o uso de predição computacional relatado em previsão de rejeição crônica de aloenxerto, função tardia do enxerto e sobrevida do enxerto a fim de trazer perspectivas sobre o assunto e abrir uma discussão sobre o futuro dessas aplicações.

Em sua pesquisa, Hannun e Andrade (2018) encontraram um consenso dentro da literatura: apesar de acertos entre alguns grupos de pacientes, ainda há a necessidade de mais pesquisas para que a aplicação seja validada e sua aplicação em uma rotina clínica seja viável, assim chegando à conclusão quanto a utilização dessas aplicações em comparação com as práticas atuais. Apesar disso, a pesquisa feita demonstra que técnicas de aprendizagem de máquinas são métodos flexíveis e viáveis para prever resultados envolvendo diversas variáveis (HANNUN; ANDRADE, 2018).

A dissertação de Silva (2017) utiliza a Mineração de Textos a fim de medir o desempenho na detecção e previsão de Eventos Adversos. Foi realizada a aplicação para prever Infecções do Sítio Cirúrgico com base em textos livres de descrições cirúrgicas no Hospital de Clínicas de Porto Alegre.

Para a segurança de pacientes, evitar e prevenir Infecções de Sítio Cirúrgico é essencial, além delas estarem entre Eventos Adversos mais comuns (SILVA, 2017). A fim de avaliar o desempenho da mineração de textos dentro desse estudo, Silva (2017) analisou 15.479 descrições de cirurgias, onde técnicas de pré-

processamento de texto e os seguintes métodos para classificação foram aplicados: Adaboost, DecisionTree, LinearSVC, Logistic Regression, Multinomial, Naive Bayes, Nearest Centroid, Random Forest, Stochastic Gradient Descent, Suport Vector Classification (SVC) (SILVA, 2017).

Em seus resultados, Silva (2017) identifica os melhores métodos de pré-processamento e mineração para prevenção de Infecções de Sítio Cirúrgico, com a possibilidade de aplicação em outros Eventos Adversos. O método Stochastic Gradient Descent obteve o melhor desempenho, 79,9 %, enquanto o método Decision Tree demonstrou o pior desempenho entre eles, 68,1 %. Direcionar ações de vigilância é uma maneira de utilizar os métodos de mineração, com o apoio das previsões de infecções (SILVA, 2017).

Feuser (2017), ex-aluno da UTFPR campus Pato Branco, aplicou o processo de KDD em prontuários eletrônicos provenientes da unidade de saúde de redes públicas, onde, através da aplicação do algoritmo Apriori e geração de regras de associação buscou encontrar conhecimentos e informações relacionados a sintomas e doenças que estão correlacionados, além de grupos de pessoas associados à algumas doenças. A base de dados utilizada é o conjunto de prontuários médicos que será estudado na próxima etapa deste trabalho, a mesma contém 43.876 registros de pacientes e 2.296.626 atendimentos.

Após tratar os dados nas fases de seleção e pré-processamento, Feuser (2017) gerou um arquivo csv onde foi possibilitada a aplicação do algoritmo Apriori, através do software WEKA. Foram utilizadas informações de: dados dos usuários, bairro e moradia de pacientes, sexo dos pacientes, informações do e-SUS e o prontuário eletrônico do paciente. Com a aplicação do algoritmo, Feuser gerou 25 regras de associação com confiança entre 92 % e 70 %, percorrendo o arquivo preparado por 18 ciclos, com 9.717 instâncias e 352 atributos (FEUSER, 2017).

Aplicado o algoritmo Apriori, Feuser (2017) analisou apenas as cinco regras com maior confiança das obtidas pelo processo, deixando as demais para a avaliação de profissionais da área da saúde. Dentre as cinco, destacam-se nas observações a de que o grupo de doenças mais comum entre os prontuários eletrônicos da Unidade de Pronto Atendimento analisada são relacionadas a infecções agudas das vias aéreas superiores; foi evidenciado que pessoas com sinais de febre de origem desconhecida estão associadas a esses casos,

contaminando pessoas que conviviam no mesmo ambiente, principalmente em escolas (FEUSER, 2017).

Por fim, Feuser (2017) conclui que a verificação das regras apresentadas poderia trazer uma expansão ao trabalho, podendo ser aplicado em outras áreas como hospitais e trazer auxílio aos profissionais da saúde e gestão de prevenção.

3.10 CONSIDERAÇÕES FINAIS

Neste Capítulo foram discutidas as tarefas realizadas antes e durante o processo de Mineração de Dados, que estão englobadas no processo de Descoberta de Conhecimento em Base de Dados, como a Limpeza de Dados, Seleção de Dados, Pré-processamento e Transformação de Dados.

Foi realizada uma introdução a conceitos de Aprendizagem de Máquinas, para que tenha-se conhecimento teórico dos algoritmos de Mineração de Dados e suas variações. Houve um foco maior em algoritmos de Regras de Associação, dos quais foram estudados o algoritmo Apriori e Hotspot.

Para finalizar o referencial teórico, foram também discutidos trabalhos relacionados, onde foram discutidos brevemente trabalhos na área da computação que possuem como tema não só a Mineração de Dados e Inteligência Artificial, como também possuem relação com a área da saúde.

4 DESENVOLVIMENTO

Nesta seção, será descrito o processo para o qual elas foram utilizadas na preparação da base de dados e execução do algoritmo de mineração de dados.

4.1 BASE DE DADOS

Os dados de prontuário eletrônico que estão contidos na base de dados foram fornecidos por um projeto onde participam a Secretaria Municipal de Saúde de Pato Branco, a Secretaria de Ciência e Tecnologia, a Universidade Tecnológica Federal do Paraná e a desenvolvedora do sistema de prontuário (IDS Desenvolvimento de Software e Assessoria ltda.). Os dados também foram utilizados para o trabalho de Rodrigo Feuser, o qual este trabalho dá continuidade.

A Base de Dados consiste de dois esquemas: *winsaude* e *publico*. O esquema *winsaude* contém trinta e três tabelas, enquanto o esquema público contém cinco tabelas. Dentro do esquema *winsaude* encontra-se o dicionário do banco de dados nas tabelas *ddcampos*, *ddtabela*, *ddcodfix*, que são tabelas utilizadas para maior compreensão dos dados exibidos na base, também facilitando o uso dos diversos campos e tabelas contidos na base.

Das tabelas contidas na base de dados, cinco delas serão usadas. A partir da Figura 7 pode-se visualizar as cinco tabelas principais desse trabalho: SASEXOS, SAUSUARI, SABAIDIS, SAFICATE, SASUSUSU. Um Modelo Entidade Relacionamento pode ser visto no Anexo A, onde estão contidos os atributos das tabelas utilizadas no trabalho.

Figura 7 - Tabelas utilizadas na base de dados

tabcodigo smallint	tabnomfis character varying (9)	tabdescri character varying (50)	tabchprim character varying (80)	tabsigla character varying (3)
108	SASEXOS	Sexos	SEXCODIGO	SEX
107	SAUSUARI	Usuários	USUCODIGO	USU
103	SABAIDIS	Bairros e Distritos	BDICODIGO	BDI
157	SAFICATE	Fichas de Atendimentos	USUCODIGO;FATCODIGO	FAT
674	SASUSUSU	Informações de e-SUS do...	USUCODIGO	ESU

Fonte: Autoria própria

Segue uma breve descrição de cada uma das cinco tabelas:

- SASEXOS: Relacionada ao sexo dos pacientes, possui três opções: masculino, feminino ou indiferente;
- SAUSUARI: Contém os usuários do sistema de saúde SUS, ou seja, tabela relacionada aos dados dos pacientes;
- SABAIDIS: Refere-se a informações de localidade dos usuários do sistema. São omitidas informações como logradouros ou ruas para que a moradia do usuário não seja descoberta, consistindo então de informações do bairro ou distrito;
- SAFICATE: Onde se encontra o prontuário eletrônico do paciente. Nela, todos os dados de fichas de atendimento são listados, onde estão contidos o processo completo da triagem de enfermeiros, medicação, alta ou encaminhamento de unidades hospitalares;
- SASUSUSU: Informações do e-SUS, como histórico de doenças e práticas rotineiras dos paciêntes, além de dados do sistema de saúde do município. É formado a partir de um questionário socioeconômico respondido por usuários.

4.2 LIMPEZA E SELEÇÃO DE DADOS

Para realizar esta etapa, foram realizadas várias *querries* no SQL, essas *querries* foram baseadas em códigos utilizados por Feuser (2017) na preparação da Base de Dados para que possa ser feita a Transformação e Mineração de Dados.

Primeiramente foi realizada a criação de uma tabela específica para o trabalho, chamada SAFICATE2. Nela, a princípio, estarão contidos todos os dados relacionados ao prontuário eletrônico. Para criá-la, utiliza-se o comando SQL *create table*. A estrutura dos atributos é igual a da tabela SAFICATE e serão inseridos todos os dados da tabela original através do comando *insert into*, descrito no Código 4.

Código 4 - Criação de tabela SAFICATE2

```
CREATE TABLE SAFICATE2 AS
SELECT * FROM SAFICATE
INSERT INTO SAFICATE2 FROM (SELECT * FROM Saficate);
```

Fonte: Autoria própria

Após a criação e inserção dos dados da tabela SAFICATE original, é feita a limpeza do atributo *cidprinci*, nas linhas onde existem registros com o valor NULL, o que significa que essas linhas representam valores onde as doenças não foram identificadas (1.639.420 registros). Similarmente, foram deletados registros que possuem o CID Z00, que representam exames gerais ou também doenças não identificadas (231.923 registros). Por fim, foram deletados também registros que estão com a situação de inativos (84.215 registros). O processo é ilustrado no Código 5.

Código 5 - Registros deletados

```
DELETE FROM winsaude.saficate2 WHERE cidprinci is null;
DELETE FROM winsaude.saficate2 WHERE cidprinci = 'Z000';
DELETE FROM winsaude.saficate2 WHERE fatsituac = 1;
```

Fonte: Autoria própria

Utilizando a seleção e contagem dos registros restantes dentro da tabela SAFICATE2 no SQL, o programa encontrou 593.135 registros de atendimentos, os quais serão utilizados para a realização da Mineração de Dados.

4.3 PRÉ-PROCESSAMENTO DOS DADOS

Devido ao elevado número de classificações de doenças dos grupos CID, foi criada uma nova coluna chamada *cidgrupos*. Com esse campo, as classificações são agrupadas por intervalos, reduzindo o número de 5404 cids para 256 grupos, que generalizam as doenças. A coluna é criada através do comando *alter table*, e de uma função que separa os cids em intervalos.

Código 6 - Trecho de função grupocid

```
DECLARE
      i integer;
      grupo varchar(5);

  BEGIN
       i = cast(substr(cidprinci,2,2) as integer);
       grupo = substr(cidprinci,1,1);

       if grupo = 'A' then
              if i < 15  then return 'A00'; end if;
              if i between 15 and 19 then return 'A15'; end if;
          if i between 20 and 29 then return 'A20'; end if;
              if i between 30 and 49 then return 'A30'; end if;
              if i between 50 and 64 then return 'A50'; end if;
              if i between 65 and 69 then return 'A65'; end if;
              if i between 70 and 74 then return 'A70'; end if;
              if i between 75 and 79 then return 'A75'; end if;
              if i between 80 and 89 then return 'A80'; end if;
              if  89 < i  then return 'A90'; end if;
        end if;
```

Fonte: FEUSER (2017)

No Código 6, é demonstrado um trecho da função *grupocid*, responsável pela seleção dos intervalos entre os grupos do campo CID, utilizando a coluna *cidprinci*. Sem a criação desses intervalos entre os campos CID, a tabela armazenaria um total de 5404 colunas com tipos de doenças CID, enquanto o PostgreSQL só oferece tabelas String com um número máximo de 1600 colunas.

Também foi criada uma função para determinar sexo dos usuários nos registros e uma função para criar faixas etárias entre os pacientes também foi criada com os mesmos objetivos, a fim de que seja mais fácil trabalhar e gerar regras, utilizando dados derivados da base original (data de nascimento, grupo cid e sexo dos pacientes).

As funções com dados derivados são executadas para que seja consistente a utilização desse tipo de dado na etapa de Mineração de Dados e, junto com a função grupocid, suas saídas resultam em tabelas muito utilizadas na saída de regras de associação do trabalho.

Código 7 - Função faixaetaria

```
DECLARE faixa CHARACTER;
  BEGIN
    SELECT
    CASE
     when FLOOR(( '2016-12-31' - public.sausuari.usudatnas )/365.25) <= 7 then
          '0'
         when FLOOR(( '2016-12-31' - public.sausuari.usudatnas )/365.25)  between 8 and 17 then
          '1'
        when  FLOOR(( '2016-12-31' - public.sausuari.usudatnas )/365.25)  between 18 and 30 then
          '2'
     when  FLOOR(( '2016-12-31' - public.sausuari.usudatnas )/365.25)  between 30 and 40 then
          '3'
     when  FLOOR(( '2016-12-31' - public.sausuari.usudatnas )/365.25)  between 40 and 60 then
          '4'
     when  FLOOR(( '2016-12-31' - public.sausuari.usudatnas )/365.25)  >= 61 then
          '5'
      END
    INTO faixa
    FROM public.sausuari
    WHERE public.sausuari.usucodigo = codigo;
    RETURN faixa;
    END;
```

Fonte: FEUSER (2017)

A função *faixaetaria* utiliza a tabela SAUSUARI para extrair o atributo (*sausuari.usudatnas*) equivalente a data de nascimento dos pacientes e subtrai o último dia do ano de 2016, quando os dados desse trabalho foram pela primeira vez gerados, assim trazendo a idade de cada paciente registrado na base de dados. Então, são divididos em faixas etárias de: menor ou igual a 7 anos, entre 8 e 17, entre 18 e 30, entre 30 e 40, entre 40 e 60 e 61 anos ou maior (ver Código 7).

Código 8 - Função tiposexo

```
 DECLARE
     sexo char;

 BEGIN

       SELECT  sextipo into sexo
          FROM sasexos,sausuari
          where sausuari.sexcodigo = sasexos.sexcodigo
                and sausuari.usucodigo = usuario;

   RETURN sexo;

 END;
```

Fonte: FEUSER (2017)

A função *tiposexo* utiliza-se de duas tabelas: *sasexos* e *sausuari*, para que a seleção de registros possa identificar qual usuário está vinculado ao registro do sexo de um paciente (ver Código 8). A partir desta função são gerados dois resultados, "*M*" para masculino e "*F*" para feminino.

4.4 TRANSFORMAÇÃO DOS DADOS

A transformação dos dados utilizados na realização deste trabalhos foi feita através de uma consulta realizada com o comando *select* do PostgreSQL, o qual permite, dentro de uma *query* e em conjunto com o comando *copy*, a exportação de dados para um arquivo no formato CSV (*Comma Separated Values,* ou Valores Separados por Vírgula) seguindo as restrições impostas pelo código e funções executadas.

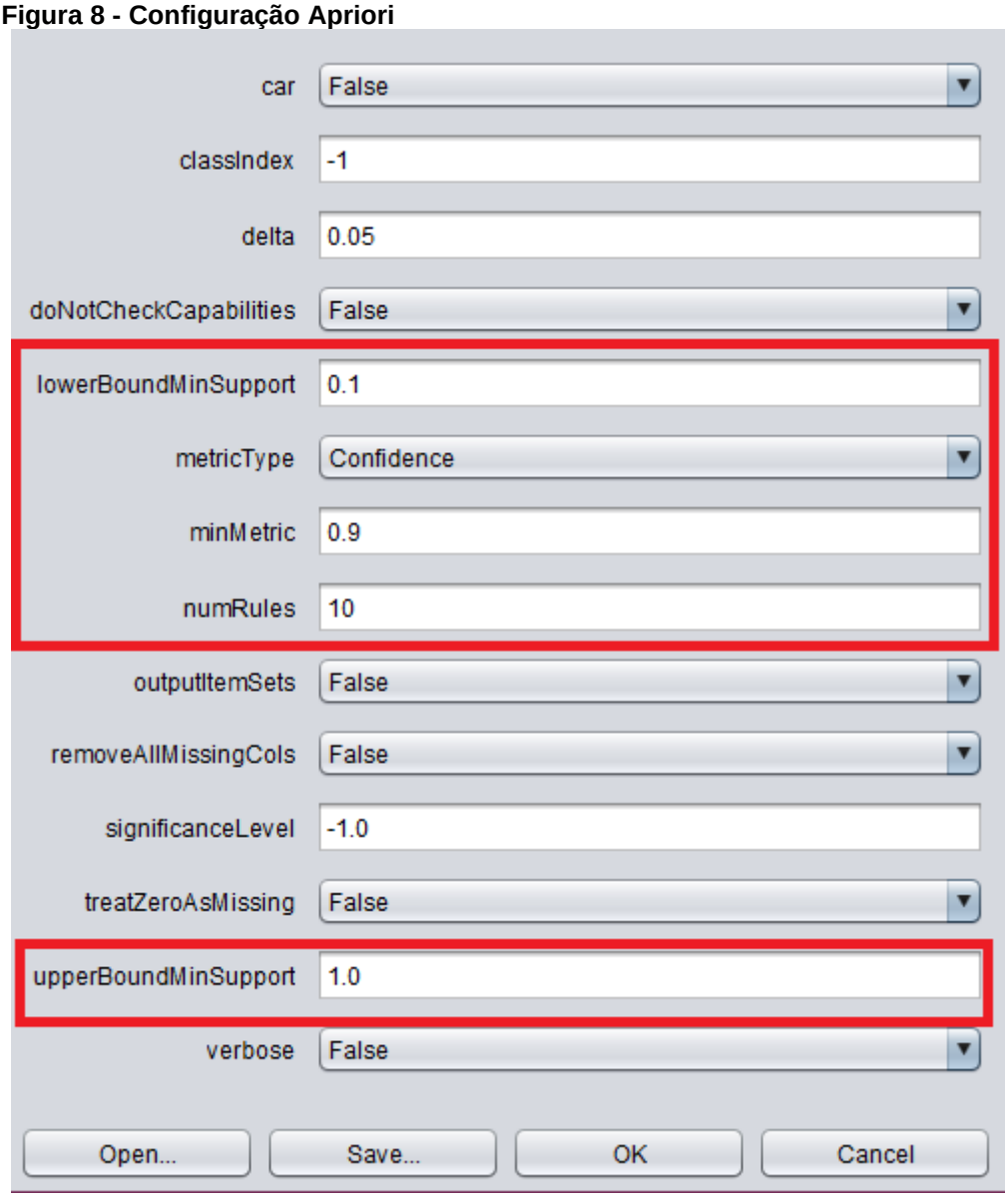
Foi possível selecionar, junto com a aplicação das funções *cidgrupos*, *faixaetaria* e *tiposexo*, todos os dados da tabela *saficate2*, separados por colunas indicando informações como o intervalo entre os grupos CID que a doença pertence, faixa etária do paciente, bairro, entre outras informações relacionadas ao atendimento. Utilizando o comando *copy* do PostgreSQL, todas essas informações foram salvas dentro de um arquivo no formato CSV, com os registros separados por vírgula, assim feita a transformação dos dados.

4.5 MINERAÇÃO DE DADOS

Com posse dos dados transformados em um arquivo CSV, os registros foram carregados dentro da Ferramenta WEKA, utilizada para execução dos algoritmos de regras de associação, que gerarão os resultados a serem estudados no trabalho e além dele.

Para a execução dos algoritmos do trabalho, foi utilizada a seção "*Associate*" localizada dentre as abas superiores na janela *Explorer* do software WEKA. Antes de executar um algoritmo, podem ser definidas diversas configurações, que altera a maneira como o algoritmo seleciona suas regras com o maior valor de confiança ou métrica selecionada, a quantia de dados desejados. Isso pode ser feito desde alterando os intervalos de suporte e confiança, no caso do algoritmo *Apriori*, até alterando a métrica pela qual o mesmo é executado.

Figura 8 - Configuração Apriori



Fonte: Autoria própria

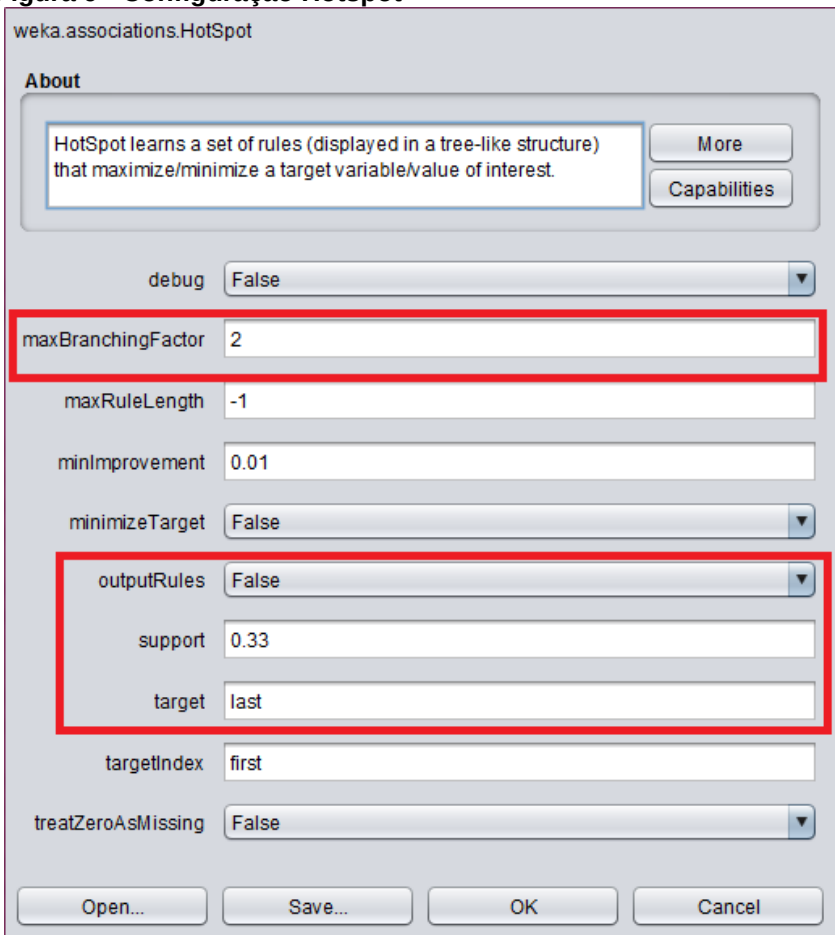
A aba de configuração do algoritmo *Apriori* pode ser vista na Figura 8. Para este trabalho, foram alteradas configurações relacionadas ao grau de suporte mínimo (*lowerBoundMinSupport*) e máximo (*upperBoundMinSupport*), confiança mínima (*minMetric*), número máximo de regras geradas (*numRules*) e métrica utilizada (*metricType*).

A configuração do Apriori com variações de confiança (*metricType* com a opção *Confidence* selecionada) a qual foi utilizada compreendia a variável de suporte mínimo (*lowerBoundMinSupport*) com o valor de 0.1, ou seja, 10%. Enquanto a confiança mínima (*minMetric*) foi variada dentre 0.1(10%) até 1.0 (100%) aumento dez porcento a cada execução do Algoritmo. Manteve-se o número de regras (*numRules*) em 100.

Para variações de Suporte, foi alterada somente a variável relacionada ao suporte mínimo (*lowerBoundMinSupport*) para 0.2 e 0.3 (20% e 30%, *respectivamente*), pois não foram encontradas mais regras para valores maiores de suporte mínimo.

Para rodar o algoritmo *Apriori* com a métrica *Lift* (*metricType* com a opção *Lift* selecionada) mostrou-se necessário alterar o valor mínimo de *Lift* para a saída do algoritmo, afim de conseguir bons resultados, já que os resultados interessantes terão pontuação maior do que 1.0. Então, decidiu alterar a variável de valor *Lift* mínimo (*minMetric*) para 1.0. Também foi estipulado um número máximo de 25 regras (*numRules*).

Figura 9 - Configuração Hotspot



Fonte: Autoria própria

Apontados na Figura 9, as configurações alteradas no algoritmo *HotSpot* para o estudo dos resultados são o valor de suporte mínimo (*support*), o atributo alvo (*target*), o número de ramos que a árvore gerará (*maxBranchingFactor*) e por fim, a

opção *outputRules,* que, quando marcada como verdadeira, traz a saída do algoritmo na forma de regras de associação.

Para este trabalho, alguns valores foram alterados para todas as instâncias onde o algoritmo *Hotspot* foi executado. São eles: o valor de suporte mínimo (*support*) foi reduzido para 0.22 (22%) afim de uma execução com uma maior saída de regras e o mesmo foi feito para o número de ramos da árvore (*maxBranchingFactor*), aumentado para 5. Por fim, a opção para trazer a saída do algoritmo como regras de associação (*outputRules*) foi selecionada como verdadeira.

Para as diferentes execuções do algoritmo foi alterado o atributo alvo (*target*), onde, após decidir os critérios para seleção de alvos, foram escolhidos: Masculino, Feminino, as diferentes faixas etárias e os grupos CID H65, M50, I10 e J00.

4.6 CONSIDERAÇÕES FINAIS

Este capítulo tratou-se dos processos executados durante a preparação dos dados, utilizando etapas de Descoberta de Conhecimento em Bases de Dados como a Seleção e Limpeza de Dados, a criação de tabelas novas a partir de dados derivados. Também foi tratada a execução do algoritmo de Mineração de Dados.

As ferramentas estudadas no capítulo 3 foram utilizadas aqui, o PostgreSQL para realizar modificações necessárias na base de dados e a ferramenta WEKA para aplicação do algoritmo de Mineração de Dados, os quais foram utilizados Apriori e Hotspot. Para o Apriori, foi feita a execução com diversos valores de confiança mínima e variações de suporte mínimo, também utilizada a métrica Lift. Para o algoritmo Hotspot, foi feito perfis de atributos específicos como sexo do paciente ou faixa etária, além de grupos CID para traçar perfil de doenças.

5 RESULTADOS

Foram realizados experimentos com dois algoritmos implementados no software WEKA: *Apriori* e *Hotspot*. Ambos são algoritmos de regras de associação e portanto podem ser feitas comparações com os seus resultados. As aplicações podem ser vistas em dois tópicos. Os experimentos com o *Apriori*, onde foram variados os valores de confiança e suporte e também aplicada a métrica *Lift*. Enquanto os experimentos com o *Hotspot* foram feitos com a busca de diversos perfis dentro da base de dados, como: sexo dos pacientes, diversas faixas etárias e também foi traçado o perfil de grupos CID específicos.

Para que seja facilitada a leitura dos resultados, foi criado também um Quadro 1, assinalando todos grupos CID evidenciados neste trabalho, que pode ser visualizado.

Quadro 1 - Informações dos grupos CID utilizados no trabalho

Número	Grupo CID	Nome
1	A00	Doenças infecciosas intestinais
2	F40	Transtornos neuróticos, relacionados com o "stress" e somatoformes
3	H65	Doenças do ouvido médio e da mastóide
4	I10	Doenças hipertensivas
5	J00	Infecções agudas das vias aéreas superiores
6	J09	Influenza [gripe] e pneumonia
7	J20	Outras infecções agudas das vias aéreas inferiores
8	M50	Outras dorsopatias
9	M70	Outros transtornos dos tecidos moles
10	Q00	Malformações congênitas do sistema nervoso
11	R10	Sintomas e sinais relativos ao aparelho digestivo e ao abdome
12	R50	Sintomas e sinais gerais
13	Z70	Pessoas em contato com os serviços de saúde em outras circunstâncias

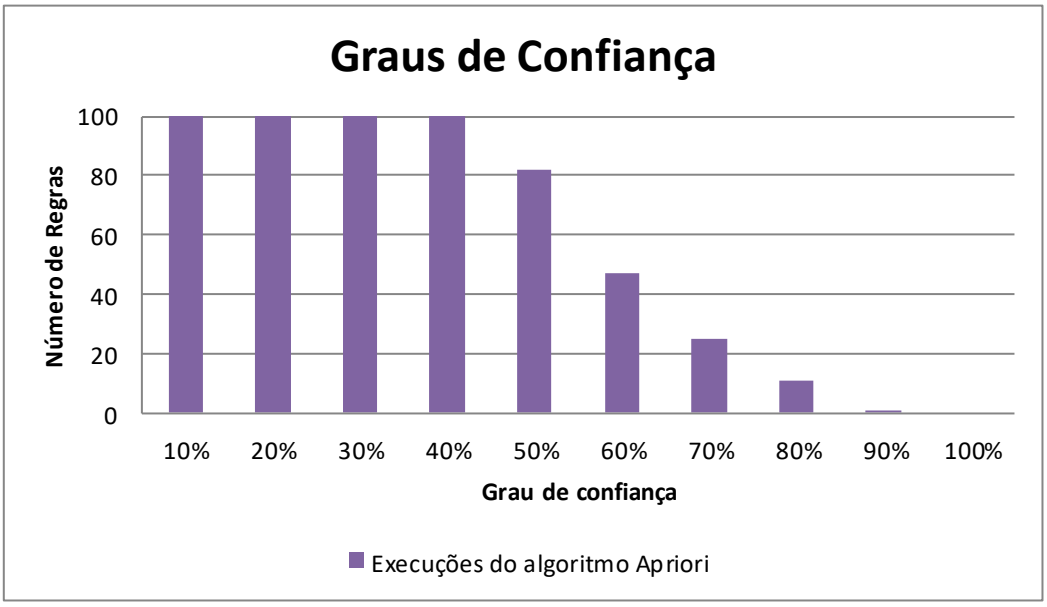
Fonte: Autoria própria

5.1 RESULTADOS APRIORI

Primeiramente, a fim de trazer alguns dados complementares ao estudo de Feuser, foi realizada a aplicação do algoritmo *Apriori*, utilizando confiança como sua métrica principal, com intervalos diferentes de níveis de confiança e suporte.

Como o nível confiança é muito baixo em regras posteriores e não seriam mais interessantes, foi imposto um limite de 100 regras para a execução do algoritmo, o que compreende os intervalos com confiança mínima de 10% a 40%. A partir dos 50% pode-se notar a redução do número de regras, até o mínimo de 90%, onde somente uma regra foi encontrada. Esse dados foram evidenciados através do Gráfico 1.

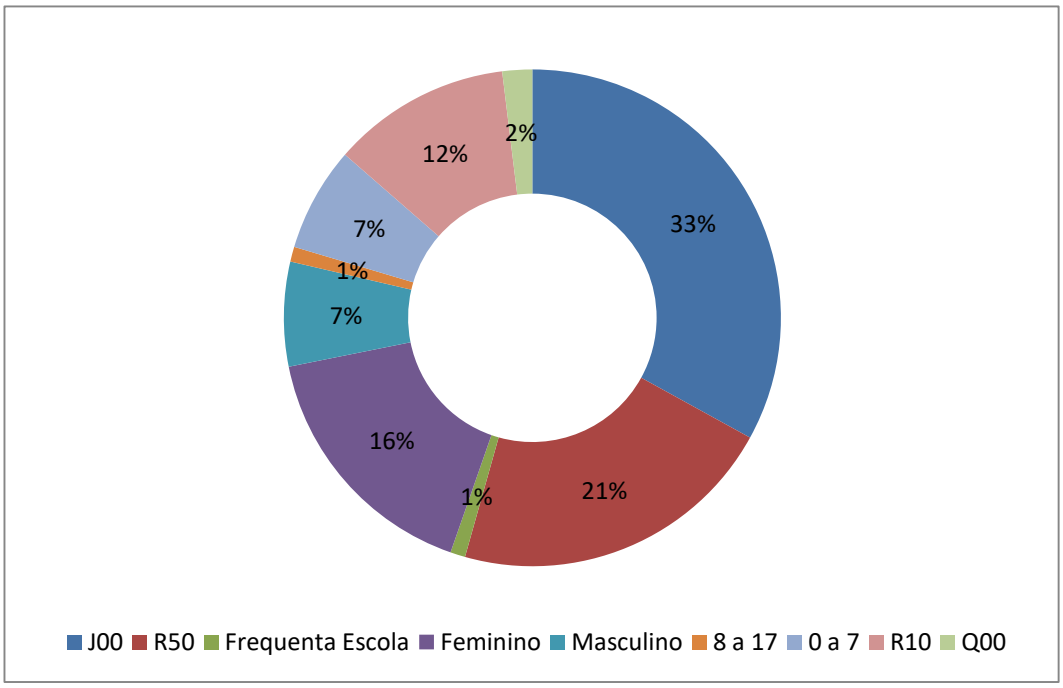
Gráfico 1 - Regras a partir do grau de confiança



Fonte: Autoria própria

Foi feita também uma comparação onde se encontram os atributos que aparecem como consequente nas 100 regras da execução do Apriori (ver Gráfico 2). Nota-se que as ocorrências mais comuns são do grupo J00 (infecções agudas das vias aéreas superiores) e R50 (sintomas e sinais gerais). Além de grupos de doenças, outros atributos que podem ser notados no gráfico são características como: se o paciente frequenta escola, as distinções de sexo masculino e feminino e também faixas de idade, mais especificamente entre 0 a 7 e 8 a 17 anos de idade.

Gráfico 2 - Atributos visualizados nas regras



Fonte: Autoria própria

Em seguida, foi feita a execução do algoritmo com valores diferentes de suporte mínimo, 20% e 30%. Relativamente diferentes, pode-se notar que o número de regras foi bastante reduzido na execução do algoritmo, restando apenas 12 regras com 20% de suporte mínimo e 4 com 30%. Não foram encontradas mais regras ao utilizar valores maiores de suporte mínimo. As regras podem ser visualizadas no Quadro 2 e Quadro 3.

Quadro 2 - Saída Apriori com Suporte 20%

Número	Premissa	Conclusão	Instâncias	Confiança
1	J20 = y	==> J00=y	1965	84%
2	0 a 7=y	==>J00=y	2569	80%
3	Q00=y	==>J00=y	2085	74%
4	R50=y	==>J00=y	3114	70%
5	R10=y	==>J00=y	2555	69%
6	Feminino=y	==>J00=y	3244	62%
7	Masculino=y	==>J00=y	2817	62%
8	R10=y	==>Feminino=y	2271	62%
9	R10=y	==>R50=y	2228	60%
10	R50=y	==>Feminino=y	2517	56%

Fonte: Autoria própria

As cinco primeiras regras denotam ocorrências do grupo CID J00, o CID mais comum dentre a base de dados e portanto o que gera mais regras. Pode-se observar na primeira regra a relação entre os grupos CID J00 e J20. Enquanto a segunda regra traz a relação entre a faixa etária de 0 a 7 anos. As próximas três regras relacionam os grupos CID Q00, R50 e R10 com o grupo J00.

As regras seis e sete mostram as ocorrências de instâncias do grupo J00 para os sexos masculino e feminino, respectivamente 2817 e 3244.

Por fim, entre as regras 8 a 10 são relacionadas as instâncias do sexo feminino nos grupos R10 e R50 e também a relação entre esses dois grupos CID. Outras duas regras foram geradas mas que repetiam informações anteriormente obtidas nas regras, com a diferença de premissa e consequente estarem invertidas e as instâncias resultantes serem as mesmas, portanto foram omitidas da apresentação de resultados.

Quadro 3 - Saída Apriori com Suporte 30%

Número	Premissa	Conclusão	Instâncias	Confiança
1	R50=y	==> J00=y	3114	70%
2	Feminino=y	==> J00=y	3244	62%

Fonte: Autoria própria

As regras geradas a partir do suporte mínimo de 30% trouxeram resultados muito similares algumas regras da execução do algoritmo com 20%, sendo dois dos

quatro resultados redundantes já que são invertidos premissa e consequente, porém as instâncias afetadas permanecem as mesmas. A regra do topo traça novamente uma relação entre o grupo R50 e a doença J00, enquanto a segunda regra denota as ocorrências da doença J00 dentre os pacientes que são do sexo feminino.

Apesar de bastante similares os resultados, as regras da saída do algoritmo estão geralmente localizadas em posições muito baixas na saída do *Apriori* normal, correlacionando alguns grupos de doença (J00 e R50 ou J00 e J20) já conhecidos da execução anterior, porém com um número reduzido de regras pode-se chegar mais rápido à definição de relações entre atributos.

Por fim, foi realizada a execução do algoritmo *Apriori* com a métrica *Lift*, que considera ambos suporte e confiança, utilizando a razão entre os dois. Para este experimento foi modificado o valor de *Lift* mínimo para 1, pois dentro da métrica *Lift* apenas são interessantes os resultados que mostrem uma pontuação maior do que 1. Para este trabalho, foi estabelecido um limite de 25 regras dentre as saídas do algoritmo, que podem ser visualizadas no Quadro 4.

Quadro 4 - Saída Apriori com métrica Lift

Número	Premissa	Conclusão	Instâncias	Confiança	Lift
1	R50=y	==>J00=y,R10=y	1676	38%	1.43
2	J00=y,R10=y	==>R50=y	1676	66%	1.43
3	R50=y	==>Q00=y	1830	41%	1.42
4	Q00=y	==>R50=y	1830	65%	1.42
5	R10=y	==>J00=y,R50=y	1676	45%	1.42
6	J00=y,R50=y	==>R10=y	1676	54%	1.42
7	J09=y	==>R50=y	1476	64%	1.39
8	R50=y	==>J09=y	1476	33%	1.39
9	J00=y	==>J20=y	1965	32%	1.35
10	J20=y	==>J00=y	1965	84%	1.35
11	R50=y	==>R10=y	2228	50%	1.32
12	R10=y	==>R50=y	2228	60%	1.32
13	0 a 7=y	==>J00=y	2569	80%	1.29
14	J00=y	==>0 a 7=y	2569	42%	1.29
15	8 a 17=y	==>J00=y	1490	79%	1.26
16	J00=y	==>8 a 17=y	1490	25%	1.26
17	A00=y	==>J00=y	1868	79%	1.26
18	J00=y	==>A00=y	1868	31%	1.26
19	J00=y	==>Frequenta Escola =y	1804	30%	1.26
20	Frequenta Escola =y	==>J00=y	1804	78%	1.26
21	R10=y	==>Feminino=y	1519	41%	1.23
22	Feminino=y,J00=y	==>R10=y	1519	47%	1.23
23	J00=y	==>R10,R50=y	1676	28%	1.21
24	R10=y,R50=y	J00=y	1676	75%	1.21
25	J00=y	Q00=y	2085	34%	1.19

Fonte: Autoria própria

Uma comparação que pode ser notada dentre as cinco primeiras regras é que os valores de confiança são mais baixos na execução do algoritmo com esta métrica, sendo a regra com o valor de *Lift* mais alto (1.43) possui apenas um valor de 38% de confiança. Essa regra mostra uma relação entre Sintomas Gerais com os grupos J00 e R10.

A regra 4 mostra a aparição do atributo Q00 (malformações congênitas no sistema nervoso), relacionado com R50 (sintomas e sinais gerais). Ela também possui um dos valores mais altos de *Lift* e uma confiança de 65%.

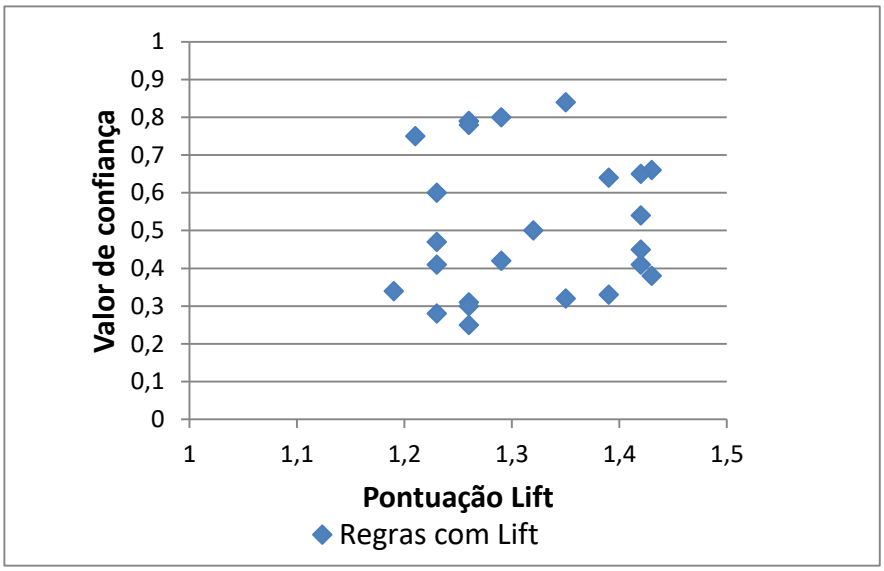
Outro grupo que pode ser enquadrado dentro os que se relacionam com sintomas e sinais gerais é o grupo J09, que representa gripe e pneumonia, na regra 8, ela possui tanto um valor de *Lift* quanto um valor confiança relativamente altos. Também pode-se notar a relação entre os grupos J00 e J20 nas regras 9 e 10.

Surgiram também regras mostrando a população de pacientes que se enquadram no grupo J00 dentre as faixas etárias de 0 a 7 anos e 8 a 17, um resultado também observado na execução do algoritmo pelo valor de confiança, especialmente considerando que ambas as regras (13 e 15) possuem um valor de 80% e 79%, respectivamente.

A relação em destaque está na regra 17, entre o atributo A00 (doenças infecciosas intestinais) e o grupo J00, com uma presença total de 1868 instâncias na execução do algoritmo.

Sobre as regras geradas pode-se, então, concluir: dentre as relações feitas entre os atributos, interessante denotar que com o valor *Lift* o algoritmo encontrou também a faixa etária de 8 a 17 anos no grupo CID J00, além de relacionar o grupo A00 (doenças infecciosas intestinais) com J00, o qual não havia aparecido durante as execuções do Apriori com confiança. Juntou-se os grupos R10 (sintomas relacionados ao aparelho digestivo e abdomen) e R50 (sintomas gerais) como possíveis causa ao grupo J00 em uma regra só. Outros atributos interessantes que aparecem relacionados são J09 (influenza e pneumonia) e J20 (infecções em vias aéreas superiores).

Gráfico 3 - Regras da execução do algoritmo com Lift



Fonte: Autoria própria

Estão representadas no Gráfico 3 as 25 regras geradas a partir da execução do algoritmo *Apriori* com *Lift*, onde o eixo X representa o valor de Lift que cada regra recebeu e o eixo Y representa o grau de confiança de cada uma delas. Pode-se perceber como várias regras, apesar de estarem com o grau de confiança baixo, permanecem em uma posição onde seu valor *Lift* é alto, o que denota as diferenças de execução das duas versões do algoritmo, podendo trazer resultados discrepantes. Assim, a diferença entre métricas pode resultar no levantamento de dados diferentes ou que podem passar despercebidos, o que para uma base de dados grande como a de atendimentos pode significar a aparição de novos atributos e regras.

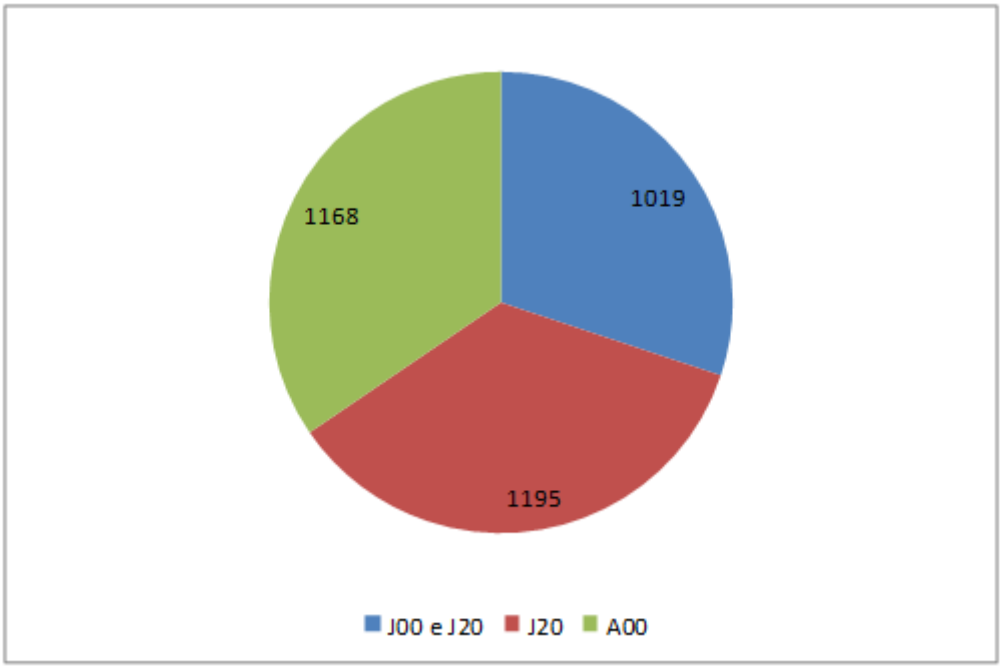
5.2 RESULTADOS HOTSPOT

Com base nas execuções do algoritmo *Apriori*, percebe-se que os grupos mais comuns são os grupos relacionados ao grupo CID J00. A fim de encontrar outras potenciais regras relacionadas a outros grupos CID, foi então decidido o uso do algoritmo *HotSpot*, que pode trazer perfis específicos considerando apenas um atributo para maximizar o seu valor de interesse. A escolha dos atributos foi feita a partir das regras do algoritmo *Apriori*, que mostrou as causas de algumas doenças dentro de grupos específicos de pacientes, diferentes de correlações entre dois grupos de doenças diferentes.

A partir deste critério, foram escolhidos, primeiramente, dois perfis para rodar o algoritmo *HotSpot*: as faixas etárias e sexo dos pacientes. Ambos aparecerem na execução do *Apriori* porém apenas relacionados com as doenças já citadas anteriormente. Busca-se, então, aprimorar os valores resultantes desses grupos de pacientes.

Executando para os pacientes de sexo Masculino, foram encontrados grupos já vistos anteriormente em regras, J00 e J20 e a maior parte das instâncias estão em uma faixa etária de 0 a 7 anos ou frequentam escola. Também foram encontradas instâncias dentro do grupo A00.

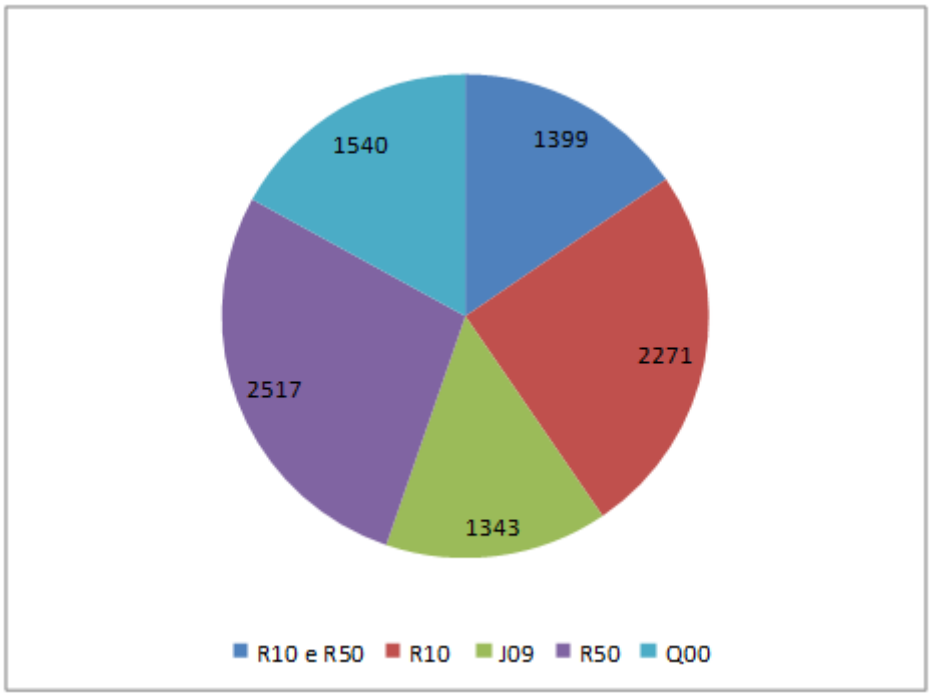
Gráfico 4 - Perfil do sexo masculino



Fonte: Autoria própria

Para o sexo feminino, os grupos mais frequentes e também correlacionados foram R10 e R50, além disso observada a ocorrência do grupo J09 e Q00, sendo um perfil discrepante da execução anterior do algoritmo em geral.

Gráfico 5 - Perfil do sexo feminino



Fonte: Autoria própria

Das informações que podem ser retiradas dos gráficos comparativos, é interessante a presença reduzida do grupo CID J00, o que pode ser notado nas instâncias do sexo feminino. Isto pode levar outros grupos que apareceram como gripe e pneumonia (J09) e malformações congênitas no sistema nervoso (Q00) como também doenças infecciosas intestinais (A00) a serem observadas com maior atenção.

Em seguida, foi realizado o experimento do algoritmo para as faixas etárias dos pacientes da base de dados. A partir das regras geradas, a Tabela 3 foi criada para assinalar os grupos CID presentes na execução do algoritmo de cada uma das faixas etárias disponíveis na base de dados.

Tabela 3 - Grupos CID por faixa etária

0 a 7	8 a 17	18 a 30	30 a 40	40 a 60	Acima de 60
H65	R10	M50	M50	M70	I10
J00	J00	R10	R50	M50	R10
J20	Q00	R50	J09	R50	Q00
A00	R50	J09		I10	M50
	A00			F40	Z70

Fonte: Autoria própria

Principalmente em idades acima de 18 anos, foram encontradas regras referentes a grupos CID que não apareceram dentre os registros e execuções do algoritmo Apriori ou apareceram em pouquíssimas regras. Logo, algumas doenças foram selecionadas para que possam ser geradas regras mais específicas por meio da execução do algoritmo *Hotspot*. Os critérios para escolha de atributos foram: os grupos CID devem aparecer em mais de uma faixa etária e não devem ter aparecido durante as execuções do algoritmo *Apriori*. Dentro desses critérios, foram escolhidos três grupos CID, são eles:

- H65: Doenças no ouvido médio e mastóide, tem sua ocorrência dentre pacientes que estão na faixa de 0 a 7 anos e também ocorre dentre algumas instâncias da faixa de 8 a 17 anos de idade;
- I10: Doenças hipertensivas, começam a aparecer em regras para pacientes da faixa etária de 40 anos para cima;
- M50: Grupo relacionado a dorsopatias: pacientes a partir da faixa de 18 anos de idade começam a se enquadrar neste grupo CID.

Para a realização deste trabalho, foram extraídas para observação regras para cada um dos três grupos de doenças:

Quadro 5 - Saída Hotspot para dorsopatias (M50)

Numero	Premissa	Conclusão	Instâncias	Confiança
1	40 a 60=y, R50=y	==>M50=y	442	59%
2	40 a 60=y, Feminino=y	==>M50=y	435	55%
3	M70=y	==>M50=y	397	54%
4	40 a 60=y	==>M50=y	677	50%
5	I10=y	==>M50=y	414	48%
6	F40=y	==>M50=y	467	47%
7	Hipertensao Arterial=y	==>M50=y	423	45%

Fonte: Autoria própria

Com as regras visualizadas no Quadro 5, o perfil do grupo CID M50 traçado foi de que a doença está correlacionada com pacientes dentre 40 a 60 e também pode se tornar uma consequência de sintomas gerais dos pacientes com essa faixa etária. Há também uma relação com o grupo M70, que faz parte do grupo de doenças relacionadas à dorsopatias.

Outras coincidências foram de que pacientes que possuem doenças hipertensivas, hipertensão arterial e também pacientes com transtornos neuróticos ou relacionados ao “stress” podem ter uma relação com o grupo das dorsopatias.

Quadro 6 - Saída Hotspot para doenças hipertensivas (I10)

Numero	Premissa	Conclusão	Instâncias	Confiança
1	Hipertensao Arterial=y, Acima de 60=y	==>I10=y	308	66%
2	Hipertensao Arterial=y, M50=y	==>I10=y	267	63%
3	Hipertensao Arterial=y	==>I10=y	574	61%
4	40 a 60=y, Hipertensao Arterial=y	==>I10=y	221	59%
5	Acima de 60=y, Feminino=y	==>I10=y	246	53%
6	Acima de 60=y, R50=y	==>I10=y	227	53%
7	Acima de 60=y	==>I10=y	410	50%
8	40 a 60=y, Feminino=y	==>I10=y	223	28%
9	Z70=y	==>I10=y	249	27%
10	M70=y	==>I10=y	197	27%

Fonte: Autoria própria

O perfil dos pacientes incluídos no grupo das doenças hipertensivas foi traçado e pode ser visualizado no Quadro 6. Percebe-se uma maioria de instâncias onde o paciente possui mais de 60 anos e hipertensão arterial. Também há uma relação da hipertensão arterial e o grupo M50 (outras dorsopatias, visto anteriormente), onde ambos resultam no grupo que está sendo analisado. Há uma maioria de pacientes do sexo feminino acima de 60 anos e também pacientes de 40 a 60 com hipertensão arterial.

Apesar da confiança baixa, pode-se observar também uma possível relação do grupo M70 (outros transtornos nos tecidos moles), que pode ser interessante visto relações anteriores do grupo analisado e o grupo M50. Também indicado com uma confiança baixa foi a regra de que pessoas do grupo Z70, que consultaram serviços de saúde em circunstâncias diferentes também podem fazer parte do grupo I10.

Quadro 7 - Saída Hotspot para doenças de ouvido e mastóide (H65)

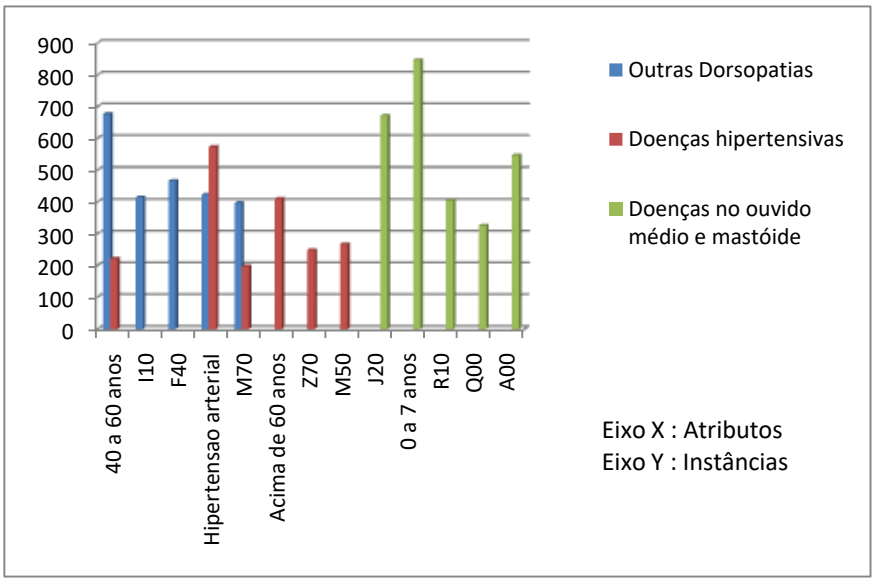
Numero	Premissa	Conclusão	Instâncias	Confiança
1	J20=y, R50=y, J00=y	==> H65=y	394	37%
2	J20=y, R10=y, J00=y	==> H65=y	347	36%
3	J20=y, A00=y	==> H65=y	321	36%
4	J20=y, R50=y	==> H65=y	408	35%
5	J20=y, Q00=y	==> H65=y	326	34%
6	J20=y, R10=y	==> H65=y	363	34%
7	J20=y, J00=y, 0 a 7=y	==> H65=y	371	33%
8	J20=y, J00=y, Masculino=y	==> H65=y	328	32%
9	0 a 7=y, R50=y	==> H65=y	331	32%
10	J20=y, J00=y	==> H65=y	621	32%

Fonte: Autoria própria

O grupo CID H65, que representa doenças do ouvido e mastóide compreende em suas regras, como pode ser visto no Quadro 7, alguns dos atributos já vistos em execuções de algoritmos anteriores e é talvez o mais próximo dos resultados originais dos experimentos. Possui, primeiramente, uma relação de consequente com três atributos, J00 (infecções agudas das vias aéreas superiores), J20 (outras infecções agudas das vias aéreas inferiores) e R50 (sintomas e sinais gerais), o que pode indicar uma conexão com as regras encontradas dentro da execução do algoritmo Apriori. Está localizado também entre pessoas com faixa etária de 0 a 7 anos com sintomas gerais e correlacionado com pacientes que possuem doenças do grupo J20 em conjunto com os grupos Q00 (malformações congênitas no sistema nervoso), R10 (sintomas e sinais relacionados ao sistema digestivo e abdome) e A00 (doenças infecciosas intestinais).

O Gráfico 6 foi feito dentre as três doenças a fim de mostrar instâncias onde ocorrem alguns atributos e grupos CID. Como pode ser notado, os grupos de dorsopatias e doenças hipertensivas têm semelhanças, compartilhando pelo menos metade dos atributos em suas regras, já o perfil do grupo de doenças no ouvido médio e mastóide não possui relação com as outras duas doenças. Ao invés disso, possui muitos atributos em comum com as execuções anteriores dos algoritmos apresentados.

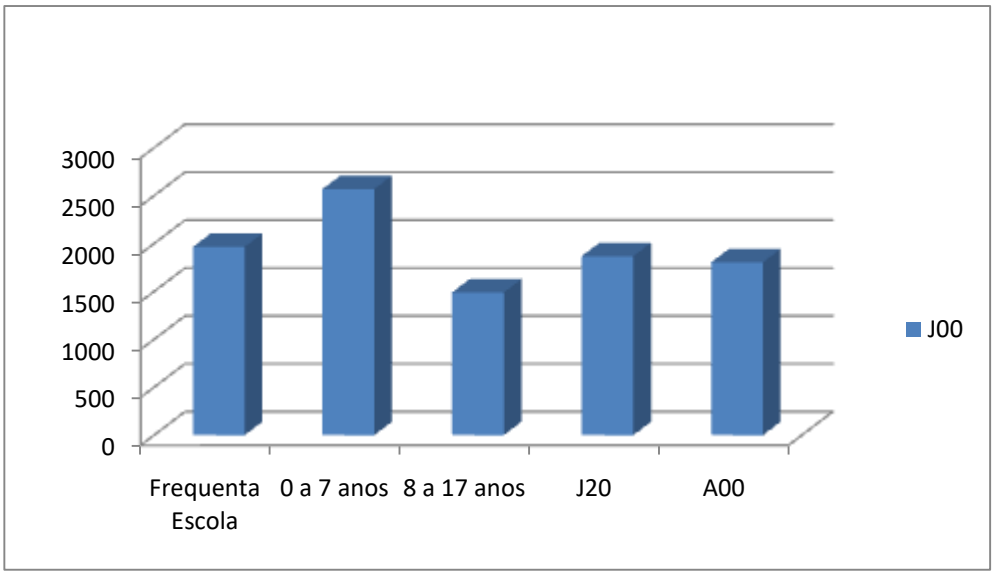
Gráfico 6 - Experimentos HotSpot com outros grupos CID



Fonte: Autoria própria

Por fim, foi traçado também o perfil do grupo de doenças J00, o qual possui maior aparecimento dentro das regras do Apriori e também em algumas execuções do HotSpot. Foi realizado, então, um gráfico referente as instâncias em que alguns atributos que apareceram na execução do algoritmo.

Gráfico 7 - Perfil grupo CID J00



Fonte: Autoria própria

O Gráfico 7 reforça as ocorrências principais de consultas que resultam no grupo CID J00, traçando algumas das principais características que definem as

instâncias onde o atributo ocorre. O maior número de pacientes está na faixa etária de 0 a 7 e isto pode ter relação com o segundo maior estar entre os pacientes que frequentam escola. Também há ocorrências, porém menores, dentro da faixa etária de 8 a 17 anos. Dos outros grupos CID, estão relacionados o grupo J20 (outras infecções agudas das vias aéreas inferiores) e o grupo A00 (doenças infecciosas intestinais).

5.3 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados os resultados dos processos realizados no desenvolvimento, das diversas aplicações dos algoritmos selecionados para o trabalho.

Para o algoritmo Apriori, foram realizados gráficos comparativos das execuções com diferentes níveis de confiança e atributos presentes nas regras de saída, também foram apresentados quadros com as regras de associação das execuções com níveis diferentes de suporte e métrica Lift. Por fim, as regras foram analisadas e foi feito outro gráfico demonstrando a comparação de pontuação Lift com nível de confiança das regras com a métrica Lift.

Para o algoritmo Hotspot foi realizado o perfil de pacientes de sexo masculino e feminino a partir de gráficos, onde são demonstrados os grupos CID provenientes das regras de saída.

Também foi criada uma tabela para visualização dos grupos CID que apareceram nas execuções do algoritmo para diferentes faixas etárias e a partir disso foram selecionados três grupos de doenças para traçar perfil: doenças no ouvido médio e mastóide, doenças hipertensivas e doenças relacionadas a dorsopatias. As regras foram então apresentadas em quadros, analisadas e foi criado um gráfico comparativo entre os três grupos.

6 CONCLUSÕES

As aplicações de Minerações de Dados mostram-se uma parte integral de estudos relacionados a conjuntos muito grandes de informações e registros atualmente, onde bases de dados podem crescer exponencialmente ao longo do tempo. Na área de saúde, isto pode ser evidenciado pois acontecem muitos atendimentos no decorrer de apenas um dia, onde cada atendimento é um registro útil em potencial.

Com as informações geradas por uma base de dados estátisticas e relatos importantes podem surgir. Os processos de Descoberta de Conhecimentos em Bases de Dados e Mineração de Dados, uma vez aplicados nestas bases de dados, podem descobrir informações úteis e relevantes dentro de quantias grandes de dados mais eficaz e rápido.

As tarefas de preparação que foram realizadas neste trabalho foram a seleção de dados, eliminação de dados inconclusivos ou não existentes e a criação de tabelas através de dados derivados. Para a Mineração de Dados, foi realizada a aplicação do algoritmo Apriori, com diversos valores de confiança e com a métrica Lift e do algoritmo Hotspot, que possibilitou a visualização de diferentes perfis de doenças.

A comparação de dados entre as execuções com valores de confiança e suporte diferentes reforçou algumas conclusões do trabalho de Feuser (2017) onde o grupo de doenças mais presente nos registro foi o de Infecções agudas das vias aéreas superiores e suas relações com pacientes que frequentam escolas e de faixas etárias entre 0 a 7 anos. Com a execução do Apriori com valores de suporte diferentes mostrou-se ainda a relação dessas doenças com Sintomas Gerais, fatores que também poderiam evidenciados com confiança, porém necessitaria de um valor baixo. Por fim, diferente de Feuser (2017), foi realizado Apriori a métrica Lift onde novos valores foram correlacionados, como o grupo de Doenças infecciosas intestinais ou a relação entre o grupo de Infecções agudas nas vias aéreas com os pacientes dentre 8 a 17.

Com a execução do algoritmo Hotspot, foi possível traçar perfis de alguns atributos presentes na base de dados utilizada no trabalho. A discrepância entre o perfil dos pacientes de sexo masculino e feminino foi, principalmente as presenças

das doenças infecciosas intestinais no masculino e malformações congênitas no sistema nervoso no feminino. Também foi realizado uma execução para todas as faixas etárias presentes nos registros e evidenciados grupos de doenças não assinalados nas execuções do Apriori. Destes, foram selecionados os grupos de dorsopatias, doenças hipertensivas e doenças no ouvido médio e mastóide.

As dorsopatias e doenças hipertensivas foram correlacionadas com pessoas de 40 a 60 anos e também pacientes hipertensão arterial, também houve a relação entre o grupo de dorsopatias com o próprio grupo de doenças hipertensivas. Por fim, ocorrências de hipertensão foram observadas em pacientes com mais de 60 anos.

Doenças no ouvido médio e mastóide mostram um perfil bastante similar ao de infecções aguas nas vias aéreas superiores, principalmente em que seus alvos se encontram dentre os pacientes de 0 a 7 anos e possui relação com os grupos de doenças infecciosas intestinais e malformações congênitas no sistema nervoso, além da aparição dos grupos de infecções agudas das vias aéreas inferiores e sintomas e sinais relativos ao abdome.

6.1 TRABALHOS FUTUROS

Para próximas pesquisas e trabalhos, as informações e dados podem então ser analisadas por um profissional da saúde e assim validadas como conhecimento útil. Pode-se também agregar outros dados relacionados à prontuários médicos, de outras unidades de saúde para realizar comparações e encontrar outras regras. Outras possibilidades interessantes seriam a de empregar outros algoritmos, não só regras de associação como também de classificação, afim de complementar e tornar cada vez mais os resultados da pesquisa interessantes.

Por fim, correlacionar os dados encontrados com outras informações e dados do e-sus. Podem ser feitas relações entre informações municipais e sociais como moradia dos pacientes, renda familiar, emprego dos pais com as doenças resultantes dos algoritmos, resultando em cada vez mais regras de associações e informações que podem ser úteis para o profissional da saúde.

REFERÊNCIAS

AMIN, Mohammad Shafenoor; CHIAM, Yin Kia; VARATHAN, Kasturi Dewi. Identification of significant features and data mining techniques in predicting heart disease. **Telematics and Informatics**. v. 36, p. 82-93, mar 2019.

DATASUS. Prontuário Eletrônico chega a 57 milhões de brasileiros. 2017. Disponível em: <http://datasus.saude.gov.br/noticias/atualizacoes/1073-prontuario-eletronico-chega-a-57-milhoes-de-brasileiros>. Acesso em: 5 nov. 2018.

DATASUS. Sistemas e Aplicativos Hospitalares. 2018. Disponível em: < http://datasus.saude.gov.br/informacoes-de-saude/62-sistemas-e-aplicativos/hospitalares>. Acesso em: 27 nov. 2018.

FARIAS, F.D.S; SOUZA, L.V.D; SOUSA, C.M; CALDAS, C.A.M; GOMES, L.F; COSTA, J.C.W.A. Data Mining Applied to Diagnose Diseases Caused by Lymphotropic Virus: a Performance Analysis. **IEEE Latin America Transactions**, v.10, n.1, p. 1319 - 1323, jan. 2012.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, Boston, v.17, n.3, p. 37-54, jul. 1996.

FEUSER. Rodrigo José. **Mineração de dados com regras de associação aplicados em dados de unidade de saúde de pronto atendimento**. 2017. 28 f. Monografia (II Curso de Especialização em Banco de Dados) - Universidade Tecnológica Federal do Paraná. Pato Branco, 2017.

GÓMEZ-VERJÁN J.C., GUTIÉRREZ-ROBLEDO L.M. The Challenge of Big Data and Data Mining in Aging Research. **Aging Research - Methodological Issues**, p. 185-196, set. 2018.

HAN, J; KAMBER, M; PEI, J. **Data Mining**: concepts and Techniques. 3. ed. Waltham: Morgan Kaufmann, 2012.

HANNUN, Pedro Guilherme Coelho; ANDRADE, Luis Gustavo Modelli de. O futuro está chegando: perspectivas promissoras sobre o uso de machine learning no transplante renal. **Brazilian Journal of Nephrology**, São Paulo, v.41, n.2, out. 2018.

HALL, Mark. Class Hotspot. 2010. Disponível em: < http://weka.sourceforge.net/doc.packages/hotSpot/weka/associations/HotSpot.html> Acesso em: 15 out. 2019.

IDS. Institucional. 2018. Disponível em: <http://www.ids.inf.br/institucional.php>. Acesso em: 10 nov. 2018.

ITI. Instituto Nacional de Tecnologia da Informação. Certificado Digital. 28 jun. 2017. Disponível em: <https://www.iti.gov.br/certificado-digital>. Acesso em: 26 nov. 2018.

KRYSZTOF, J. C; MOORE, G. W. Uniqueness of Medical Data Mining. **Artificial Intelligence in Medicine**, v.26, p 1-24, 2002.

MARIN, H.F; MASSAD, E; NETO, R.S.A. **O Prontuário Eletrônico do Paciente na assistência, informação e conhecimento médico**. São Paulo, 2003.

MAIMON, O; ROKACH, L. **The Data Mining and Knowledge Discovery Handbook**: a Complete Guide for Practitioners and Researchers. 2. ed. 2010.

MOREIRA, Leonard Barreto; NAMEN, Anderson Amendoeira. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. **Computer Methods and Programs in Biomedicine,** v. 165, p. 139-149, out. 2018.

NETTO, J.A; CAMARGO, J.T.F. Winsaude: um software de apoio aos serviços de saúde. **Interciência & Sociedade**, v.2, n.2, 2013.

PATRÍCIO, C. M. et al. O prontuário eletrônico do paciente no sistema de saúde brasileiro: uma realidade para os médicos? **Scientia Medica**, Porto Alegre, v. 21, n. 3, 2011.

PENSESUS. SUS: O que é? 2018. Disponível em: < https://pensesus.fiocruz.br/sus>. Acesso em: 10 nov. 2018.

PENTAHO. Hotspot Segmentation Profiling. 24 ago. 2010. Disponível em: < http://weka.sourceforge.net/doc.packages/hotSpot/weka/associations/HotSpot.html> Acesso em: 15 out. 2019.

PORTALDAB. O que é e-SUS AB. 2018. Disponível em: <http://dab.saude.gov.br/portaldab/o_que_e_esus_ab.php>. Acesso em: 10 nov. 2018.

PORTALDAB. Documentos. 2018. Disponível em: < http://dab.saude.gov.br/portaldab/esus.php?conteudo=documentos>. Acesso em: 15 nov. 2018.

PRASS, F. S. KDD – Uma visão geral do processo. Jul. 2012. Disponível em: <http://fp2.com.br/blog/wp-content/uploads/2012/07/KDD_Uma_visao_geral_do_processo.pdf>. Acesso em: 15 nov. 2018.

ROSELLA. Hotspot Analysis and Hotspot Software Tools. 2005. Disponível em: <http://www.roselladb.com/hot-spot.htm> Acesso em: 18 out. 2019.

SANTOS, F. Breve história dos registros hospitalares. 28 ago. 2007. Disponível em: <http://osnobresescritores.blogspot.com/2007/08/breve-histria-dos-registros.html>. Acesso em: 5 nov. 2018.

SITTIG, D.F. Advantages of computer-based medical records. **Informatics Review**. 1999.

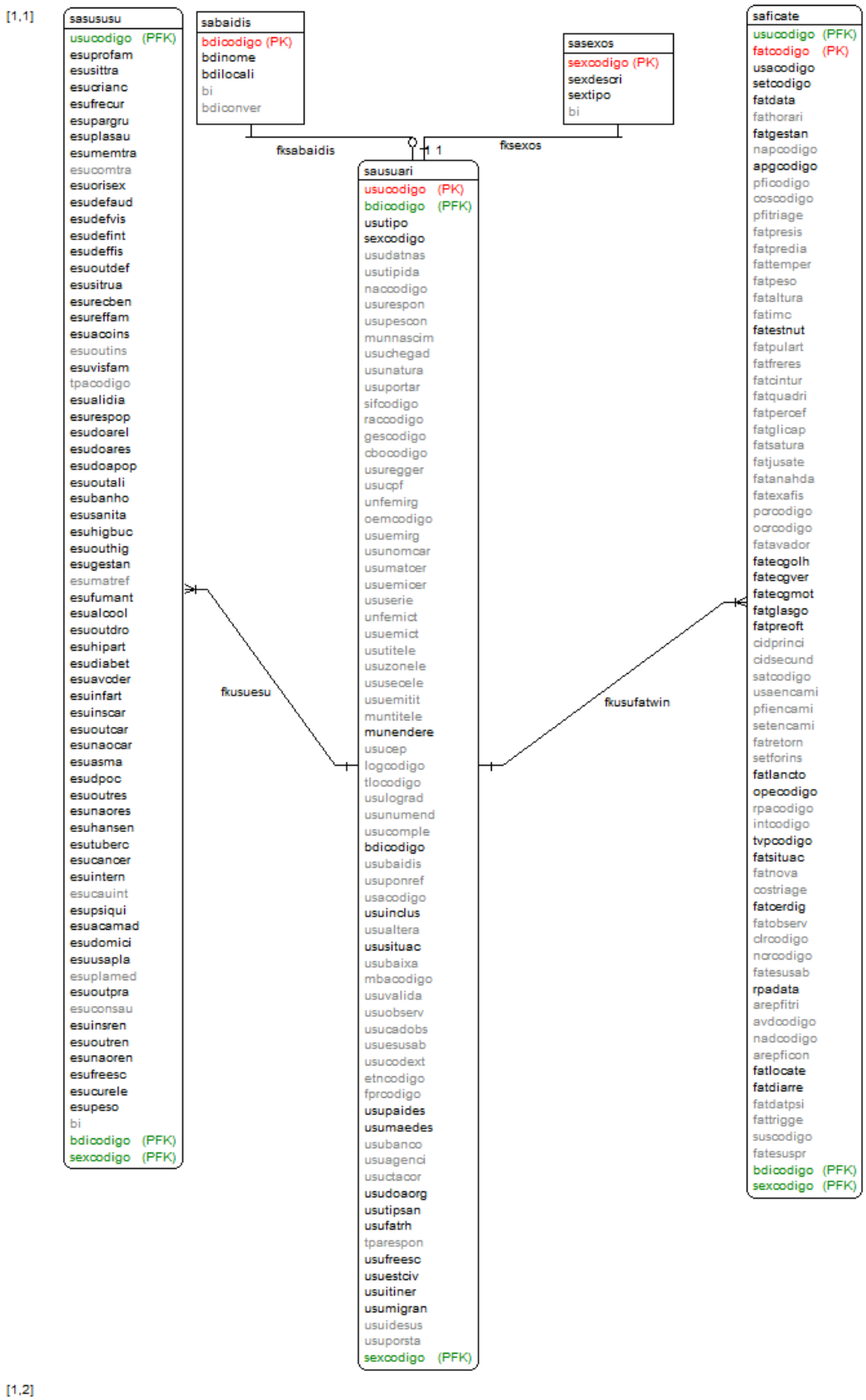
SILVA, Daniel Antonio da. **Mineração de textos aplicada na previsão e detecção de eventos adversos no Hospital de Clínicas de Porto Alegre.** 2017. 99 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Engenharia de Produção, Universa Tecnologia Federal do Rio Grande do Sul. Porto Alegre, 2017.

TAN, P; STEINBACK M; KUMAR V. ***Introduction to Data Mining***. 2. ed. Pearson Addison-Wesley. Boston, 2009.

WEKA. Wakaito Environment for Knowledge Analysis. Disponível em: <https://www.cs.waikato.ac.nz/ml/index.html> Acesso em: 3 out. 2018.

WITTEN, I.H; FRANK, E; HALL, M. A. **Data Mining**: practical Machine Learning Tools and Techniques. 3. ed. 2011.

ANEXO A - Modelo Entidade Relacionamento.



Fonte: Feuser (2017)