UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DEPARTAMENTO ACADÊMICO DE INFORMÁTICA

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

KAIQUE AUGUSTO MORAES DA SILVA

ANÁLISE DE PERFIS DE DOENÇAS COM BASE EM TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA

2019

KAIQUE AUGUSTO MORAES DA SILVA

ANÁLISE DE PERFIS DE DOENÇAS COM BASE EM TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Trabalho de Conclusão de Curso apresentada como requisito parcial à obtenção do título de Bacharel em Ciência da Computação, do Departamento Acadêmico 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
Campus Ponta Grossa

Nome da Diretoria
Nome da Coordenação
Nome do Curso

TERMO DE APROVAÇÃO

ANÁLISE DE PERFIS DE DOENÇAS COM BASE EM TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

por

KAIQUE AUGUSTO MORAES DA SILVA

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 10 de junho de 2019 como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação. O candidato foi arguido(a) 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
Membro titular

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

_____
Prof(a). MSc. Saulo Jorge Beltrão de Queiroz
Coordenador do curso

Dedico este trabalho à minha família.

AGRADECIMENTOS

Agradeço à Deus por me conceder saúde, força e perseverança para concluir mais esta etapa da minha vida. Ele que nos dá força nos momentos difíceis e nos faz acreditar que nossos sonhos são possíveis.

A minha noiva Natália, que esteve sempre ao meu lado mesmo nos momentos difíceis, me apoiando, me incentivando e acreditando em mim.

A minha mãe Andrea e ao meu padrasto Caio que me proporcionaram tudo para que eu pudesse chegar aqui.

Ao meu professor orientador Prof. Dr. André Pinz Borges, pela sabedoria e dedicação com a qual me guiou em suas orientações e revisões.

Por fim, agradeço a todos que contribuíram de alguma forma a concluir este trabalho.

“Estabelecer meta é o primeiro passo
para transformar o que é invisível visível”.
(Tony Robbins)

RESUMO

SILVA, KAIQUE. **ANÁLISE DE PERFIS DE DOENÇAS COM BASE EM TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS**. 2019. Trabalho de Conclusão de Curso Bacharelado em Ciência da Computação - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2019.

Este trabalho tem como objetivo fazer análises de perfis de doenças com base em técnicas de descoberta do conhecimento de bases de dados do prontuário eletrônico do paciente da rede pública. A descoberta de informações correlacionadas em prontuários pode ser uma tarefa complexa e de longa duração para profissionais da área da saúde, principalmente quando a quantidade de dados é de grande porte ou quando sem auxílio de sistemas computacionais especializados. O processo de descoberta do conhecimento pode ser uma alternativa para obtenção de um conhecimento especializado a partir da interpretação dos dados e transformá-los em conhecimento útil na área da saúde. Para isto, foi utilizado um banco de dados de prontuários eletrônicos de usuários do sistema único de saúde que possui 43.879 pacientes e 2.296.626 atendimentos feitos no ano de 2015 no município de Pato Branco, Paraná. Neste trabalho foi feito a descoberta de conhecimento afim de encontrar perfis de doenças utilizando as ferramentas PostgreSQL para utilização do banco de dados, os algoritmos j48, *bagging* e *boosting*, com a ferramenta WEKA. Todas tecnologias empregadas são de utilização pública. O resultado foi a obtenção e dois grupos de doenças, neoplasias que possuem 119 instâncias e traumatismos que por sua vez possui 236 instâncias, para a aplicação dos algoritmos de mineração de dados com uma seleção de 17 atributos totais. Onde o melhor desempenho foi com a execução do algoritmo *Boosting* para classificar. As regras geradas pela árvore de decisão sobre as doenças do grupo de neoplasias tiveram principais ramos bairro, peso, frequência escolar, faixa etária e altura. Já para o grupo de doenças de traumatismos as principais regras foram sobre bairro, peso, altura, faixa etária, sexo, diabetes e se fumavam. Com base nos resultados é possível fazer uma análise de um especialista da área da saúde para confirmar essas regras geradas possibilitam uma nova obtenção de conhecimento relacionado ao perfil dessas doenças para que seja possível ter uma maior prevenção e concluído a descoberta de conhecimento.

Palavras-chave: Mineração de dados. Descoberta de conhecimento. Prontuários eletrônicos. C4.5. Bagging. Boosting

ABSTRACT

SILVA, KAIQUE. **Knowledge Discovery in Databases of Eletronic Medical Records.** 2019. 24f. Work of Conclusion Course Graduation in Computer Science - Federal Technology University – Paraná. Ponta Grossa, 2019.

This study aims to analyze disease profiles based on techniques for discovering the knowledge of databases of patients' medical records in the public network. The discovery of correlated information in medical records can be a complex and long-term task for health professionals, especially when the amount of data is large or without the aid of specialized computer systems. The process of knowledge discovery can be an alternative to obtain specialized knowledge from the interpretation of data and transform it into useful knowledge in the health area. For this purpose, a database of electronic medical records of users of the single health system was used, which has 43,879 patients and 2,296,626 patients in the city of Pato Branco, Paraná, in the year 2015. In this work, knowledge was discovered in order to find disease profiles using PostgreSQL tools for database use, j48, bagging and boosting algorithms with the WEKA tool. All technologies employed are of public use. The result was the acquisition and two groups of diseases, neoplasms that have 119 instances and trauma that in turn have 236 instances, for the application of the algorithms of data mining with a selection of 17 total attributes. Where the best performance was with running the Boosting algorithm to sort. The rules generated by the decision tree on the diseases of the group of neoplasms had main neighborhood branches, weight, school attendance, age group and height. Already for the group of diseases of injuries the main rules were about neighborhood, weight, height, age group, sex, diabetes and if they smoked. Based on the results it is possible to make an analysis of a specialist in the health area to confirm that the rules generated enable a new knowledge acquisition related to the profile of these diseases so that it is possible to have a greater prevention and complete the discovery of knowledge.

Keywords: Data mining. Knowledge discovery. Electronic medical record. C4.5, Bagging, Boosting.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

ARFF	*Attribute-Relation File Format*
ASCII	*American Standard Code for Information Interchance*
CID-10	Classificação Internacional de Doenças - Décima Revisão
CSV	*Comma-Separrated Value*
CPF	Cadastro de Pessoas Físicas
CNES	Cadastro Nacional dos Estabelecimentos de Saúde
IDS	Desenvolvimento de Software e Assessoria Ltda
IOM	*Institute of Medice*
KDD	*Knowledge Discovery in Databases*
MD	Mineração de Dados
PEP	Prontuário Eletrônico do Paciente
PHP	PHP: Hypertext Preprocessor
SGBD	Sistema Gerenciador de Banco de Dados
SUS	Sistema Único de Saúde
UPA	Unidade de Pronto Atendimento
UBS	Unidade Básica de Saúde
UTFPR	Universidade Tecnológica Federal do Paraná
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Bancos de dados são comumente utilizados para arquivar dados de empresas dos mais diversos ramos, como consultórios e hospitais (WITTEN, 2011). Grande parte destes dados são armazenados e não são explorados pelas empresas, as quais necessitam de melhorias nestas áreas. Algumas dessas empresas, como hospitais, possuem um modo específico para armazenamento de dados, os prontuários eletrônicos (WITTEN, 2011).

Prontuários eletrônicos são formulários que foram preenchidos durante o atendimento ou consulta de uma pessoa na área da saúde, estes dados podem ajudar o médico a descobrir uma possível causa de uma doença. Os prontuários também são preenchidos após a consulta médica sendo colocados os diagnósticos desta pessoa.

Os prontuários normalmente possuem dados como nome, idade, profissão, problemas de saúde do paciente, entre outros. O uso destes tem ajudado cada vez mais as entidades de saúde a gerenciar dados, informações destes pacientes, profissionais, médicos etc. (KRYSZTOF, 2002).

A busca por padrões em dados existe a muito tempo e já teve outros nomes, como extração de conhecimento, descoberta de informação, processamento padronizado em dados e mineração de dados (FAYYAD, 1996).

Um dos meios para explorar estes dados de maneira útil é por meio da Descoberta de Conhecimento em Banco de Dados (em inglês, *Knowledge Discovery in Databases* - KDD). Por KDD entende-se o processo, normalmente não trivial, de obter informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados (FAYYAD, 1996). KDD permite obter novas informações após uma série de etapas: seleção dos dados, o pré-processamento e limpeza destes, transformação, mineração de dados e interpretação/avaliação dos resultados obtidos.

O processo de KDD é iterativo e interativo, este envolve alguns passos com muitas escolhas que são feitas pelo usuário. (FAYYAD, 1996). Fayyad também fala que a descoberta de conhecimento tem como objetivos a verificação e descoberta. A verificação é a verificação das hipóteses do usuário para o sistema. Com descoberta pode-se encontrar novos padrões, que se divide em predição, onde o sistema

descobre padrões de previsão futuros, e descrição, onde o sistema descobre padrões para representação compreensível para o ser humano.

As atividades nas empresas, como clínicas, auxiliam na aquisição de grandes quantidades de dados os quais são armazenados separadamente em um prontuário individual (GALVÃO, 2009).

Mesmo com os sistemas computacionais, não é fácil a obtenção do controle e ordem de todos estes dados gerados. O KDD fornece uma possibilidade de adquirir esse conhecimento que normalmente é implícito inerente a um relacionamento matemático (GALVÃO, 2009).

Segundo John Halamka (2006), médico especializado em prontuários eletrônicos e tratamento de dados, a tecnologia de informação em saúde possui pontos positivos quando implantados nos serviços de saúde, tais como:

- Aumento da adesão aos protocolos clínicos;
- Incremento da capacidade de executar a vigilância e monitorar condições da doença;
- Redução de erros nas medicações e custos de diagnósticos; e
- Maior aproveitamento do tempo dos profissionais de saúde.

Há uma apreensão por parte de gestores e profissionais de saúde na hora de compreender os dados e utilizar a informação e conhecimento obtidos para promover uma gestão da informação e qualidade. Isso é decorrência do ritmo acelerado de geração de dados do banco, que produz uma um grande volume de dados que é natural o humano não conseguir explorar, extrair e interpretar estes mesmos dados sem ajuda (GALVÃO, 2009).

Neste trabalho são aplicadas etapas do KDD e algoritmos de Mineração de Dados em um conjunto de dados de prontuários eletrônicos de usuários do Sistema Único de Saúde (SUS) com o intuito de analisar perfis de doenças. O conjunto de dados compreende 43.879 pacientes registrados e 2.296.626 atendimentos ocorridos no município de Pato Branco, Paraná no período de 2015 a 2016.

1.1 OBJETIVOS

O trabalho possui os objetivos gerais estabelecidos na sessão 1.1.1 e os objetivos específicos na sessão 1.1.2.

1.1.1 Objetivo Geral

Este trabalho tem como objetivo aplicar as etapas do método KDD em uma base de dados de prontuários eletrônicos de usuários do Sistema Único de Saúde (SUS), a fim de descobrir padrões de informações sobre pacientes que possuem determinadas doenças.

1.1.2 Objetivo Específicos

Para corroborar com o objetivo geral, foram definidos os seguintes objetivos específicos:

- Analisar o banco de dados de prontuários médicos obtidos;
- Levantar, na literatura, trabalhos similares para determinar e analisar algoritmos similares aplicáveis no mesmo domínio de problema;
- Determinar os algoritmos aplicáveis e aplicá-los no banco de dados transformado;
- Aplicar as etapas de KDD para obtenção de conhecimento a partir do conjunto de dados e identificação de possíveis grupos de doença para mineração;
- Gerar, por meio da mineração de dados, modelos sobre alguns grupos de doenças;
- Visualizar e analisar os resultados obtidos.

1.2 JUSTIFICATIVA

Com o KDD é possível manipular os dados e selecioná-los de acordo com as necessidades, gerando assim dados de interesse os quais podem ser transformados

em conhecimento por meio da visualização do resultado da mineração de dados. Se empregada em programas de prontuários eletrônicos pode auxiliar na identificação das enfermidades ou padrões de perfis de usuários, para obtenção de um sistema preventivo para auxiliar o tratamento destes.

O KDD possibilita que pessoas da área da tecnologia possam ajudar a descobrir conhecimentos relevantes sobre outras áreas. Durante a fase de análise dos resultados analistas conseguem olhar para menos dados podendo ter uma análise mais especifica para identificação de um novo conhecimento.

A utilização da mineração de dados para realizar descoberta de conhecimento em banco de dados de prontuários médicos geram regras de associação para identificar qual os maiores índices de doenças em certas faixas etárias, como também os períodos com maiores ocorrências de determinadas enfermidades, garantindo assim um sistema de prevenção mais eficaz nas unidades de pronto atendimento (FEUSER, 2017).

1.3 ESCOPO DO TRABALHO

Este trabalho apresenta as seguintes limitações: (i) será utilizado algoritmos de Mineração de Dados tradicionais para descoberta de conhecimento; (ii) não serão desenvolvidos novos algoritmos e técnicas de descoberta de conhecimento; (iii) os resultados obtidos foram comparados em termos técnicos, como percentuais de acerto dos modelos, devido à dificuldade de contato com profissionais da área da saúde.

1.4 ORGANIZAÇÃO DO TRABALHO

O presente trabalho está organizado da seguinte forma: no Capítulo 2 serão apresentados o referencial teórico com os conceitos e técnicas referentes à Descoberta de Conhecimento em Bases de Dados, o prontuário eletrônico e trabalhos relacionados. No Capítulo 3 será apresentado o desenvolvimento da base de dados e as etapas do KDD. O Capítulo 4 apresenta detalhes sobre os resultados usando os algoritmos J48, *Bagging* e *Boosting* com J48. O Capítulo 5 apresenta a conclusão do trabalho, assim como suas limitações e possíveis trabalhos futuros.

2 REFERENCIAL TEÓRICO

Tem-se uma urgência para a geração de novas teorias computacionais para ajudar-nos a extrair informações úteis do alto crescimento dos volumes de dados digitais (FAYYAD, 1996).

As pessoas são sobrecarregadas com informações. O montante de dados no mundo e na vida das pessoas cresce cada vez mais. Computadores, celulares com cada vez mais memória e meios para salvar dados online, como nuvens, fazem com que coisas que antes poderiam ser facilmente apagadas sejam guardadas (WITTEN, 2011).

A mineração de dados consiste em resolver problemas analisando dados já existentes em nossos bancos de dados (WITTEN, 2011).

2.1 BASE DE DADOS

Para o desenvolvimento deste trabalho é utilizado o gerenciador de banco de dados objeto relacional (SGBD) chamado *PostgreSQL,* desenvolvido como projeto de código aberto.

PostgreSQL consegue suportar todos os tipos de dados que são usados atualmente como documentos *json*, *xml*, chaves primarias, índices para tabela de partição entre outros. Este consegue atender todos os requisitos para que possa ser feito o trabalho, tais como:

- Recuperar um arquivo backup dos dados;
- Criar tabelas com os mesmos atributos da tabela principal a tabela de dados do banco recuperado;
- Inserir dados em tabelas para aplicação das etapas do KDD;
- Filtrar e apagar dados dessa tabela com uso dos comandos *delete* entre outros;
- Exportar esta tabela após aplicada as etapas possíveis do KDD nela para um arquivo CSV.

Com a ferramenta serão utilizados comandos do SQL para seleção, alteração, inserção, remoção, criação de *views* e transformação dos dados, tais funcionalidades

tornam possível a aplicação de descoberta de conhecimento na base de dados garantindo a execução das etapas que antecedem a mineração de dados.

2.2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Dados em estado bruto raramente oferecem benefícios diretos para processo de tomada de decisão. Tradicionalmente, é necessária uma análise destes dados, normalmente restrita a um processo manual, na qual analistas devem estar familiarizados. Tal fato gera um grande problema pois, quando os dados possuem um incremento constante, a análise pode se tornar lenta e difícil de ser executada (FELIX, 1998).

O processo tipicamente utilizado para realizar as análises de dados fez com que pesquisadores demonstrassem interesse em analisar informações. Este interesse resultou na área denominada atualmente como Descoberta de Conhecimento em Bases de Dados (em inglês, *Knowledge Discovery in Databases* – KDD) (WITTEN, 2011).

O KDD tem como premissa transformar um alto volume de dados em um conjunto menor dele com ênfase em retirar dados que não são interessantes para o estudo e deixar apenas os dados que possuam conhecimento agregado. Este conhecimento pode vir de formas diversas de representação como gráficos e matrizes de confusão dependendo, na maioria das vezes, da forma como estes dados foram obtidos e minerados.

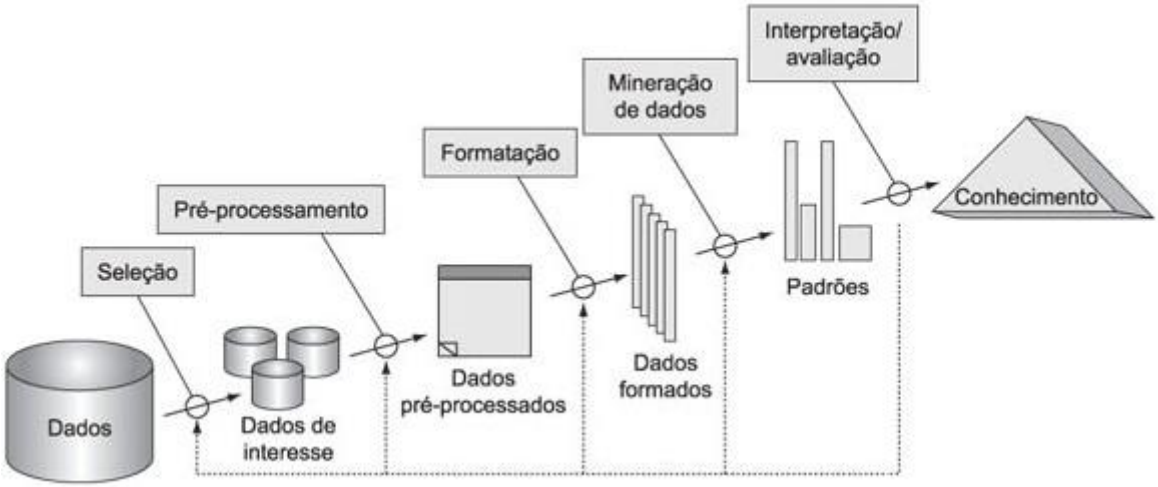
A Descoberta de Conhecimento em Bases de Dados (KDD) coloca uma ênfase especial na procura de padrões que podem ser para conhecimento de interesse ou útil, um exemplo seria as redes neurais (FAYYAD, 1996). Segundo Dias (2002) diversas áreas do conhecimento já utilizam técnicas de mineração de dados, destacando-se: o Marketing, que faz o uso de técnicas de mineração de dados para descobrir preferências do consumidor, padrões de compras para segmentação e personalização de programas de marketing; Finanças, que usa técnicas para descobrir padrões de fraude e desenvolver modelos de previsão e de inadimplência e por fim o Controle de processos e de qualidade, o qual utiliza algumas técnicas para auxiliar no planejamento estratégico de linhas de produção e na busca por padrões de condições físicas na embalagem e no armazenamento de produtos.

A técnica de Mineração de Dados pode também ser usada na Medicina para caracterizar comportamentos de pacientes, identificar terapias médicas de sucesso para diferentes doenças, buscar por padrões de novas doenças e para previsão de epidemias. A aplicação do KDD área de prontuários possui trabalhos relacionados, Feuser (2017) em seu trabalho realizou a classificação com o algoritmo apriori para prontuários eletrônicos.

2.2.1 Etapas do KDD

O KDD em seu processo utiliza conceitos de base de dados, métodos estatísticos, ferramentas para visualização e técnicas de Inteligência Artificial (IA). O processo divide-se nas etapas de seleção, pré-processamento, transformação, Mineração de Dados (MD) e interpretação de dados, tais etapas são indicadas na Figura 1.

Figura 1 - Etapas do processo de KDD



Fonte: Adaptado de (Fayyad, 1996)

Na primeira etapa, seleção, é feita a escolha dos dados de interesse para o KDD. Após a seleção é possível começar o pré-processamento onde os dados são alterados para que sejam filtrados ou removidos para formatação posterior. Na etapa de formatação é feita a transformação dos dados que restaram para que seja possível a mineração dos dados. Já na mineração dos dados algoritmos são aplicados de modo a classificar os dados gerando padrões, os quais podem ser avaliados para a descoberta de conhecimento sobre os dados de interesse. Durante o processo de

interpretação é possível voltar as etapas anteriores para que seja melhorado a saída e obtenha-se melhores resultados. Ao completar o processo obtemos conhecimento sobre o banco de dados que foi usado para este processo.

2.2.2 Seleção de Dados

A fase de seleção dos dados é a primeira no processo de descobrimento de conhecimento. Nesta fase é escolhido o conjunto de dados, pertencente à um domínio, contendo todas as possíveis variáveis (também chamadas de características ou atributos) e registros (também chamados de casos ou observações) que farão parte da análise. Normalmente a escolha dos dados fica a critério de um especialista do domínio, uma vez que as características dos dados podem variar de domínio para domínio (WITTEN, 2011).

O processo de seleção é complexo, uma vez que os dados podem vir de uma série de fontes diferentes (*data warehouses*, planilhas, sistemas legados) e podem possuir os mais diversos formatos. Este passo impacta a qualidade do resultado do processo, pois o conhecimento gerado será baseado nos dados selecionados nesta fase.

2.2.3 Pré-processamento e limpeza dos dados

No pré-processamento e limpeza dos dados são realizadas etapas as quais visam eliminar dados redundantes e inconsistentes, recuperar dados incompletos e avaliar possíveis dados discrepantes ao conjunto (*outliers*). Estas etapas têm como objetivo deixar o modelo o mais consistente possível que seja feita a transformação dos dados e posterior mineração.

Nesta fase é preparado o *input,* ou seja, a entrada de dados para a próxima etapa, mineração. Normalmente está fase consome tempo junto a fase de pré-processamento (WITTEN, 2011).

Também são utilizados métodos de redução ou transformação para diminuir o número de variáveis envolvidas no processo visando, com isto melhorar, o desempenho do algoritmo de análise. Estes métodos são utilizados de modo a tornar

os dados concisos para uso posterior na fase de formatação e transformação de dados.

2.2.3.1 Dados ausentes

Um problema bastante comum no pré-processamento de dados é a ausência de valores para determinadas variáveis (FAYYAD, 1996). Em outras palavras, registros com dados incompletos, seja por falhas no processo de seleção ou de revisão. O tratamento destes casos é necessário para que os resultados do processo de mineração sejam confiáveis. Existem basicamente três alternativas de solução para esse problema: usar técnicas de imputação, substituir o valor faltante pela média aritmética da variável ou excluir o registro inteiro.

2.2.3.2 Dados discrepantes

Dados que possuem valores extremos, atípicos ou com características bastante distintas dos demais registros são chamados de discrepantes, ou *outliers* (WITTEN, 2011). Estes dados podem ser atributos como uma idade incomum, ou uma altura fora de um padrão. Um exemplo para dados discrepantes seria um idoso(a) com 150 anos, visto que se sabe que está é uma idade não atingida por um ser humano. Logo, este dado é um dado discrepante.

Normalmente, registros que contêm valores outliers podem ser descartados da amostra ou alterados para a média aritmética, uma vez que estes valores alteram o resultado da mineração. Porém isto só deve ocorrer quando o dado *outlier* representar um erro de observação, de medida ou algum outro problema similar.

Antes da remoção o dado deve ser cuidadosamente analisado, pois embora atípico, o valor pode representar um dado verdadeiro. Tal análise é feita por um analista do banco de dados que deve escolher quais atributos atípicos para estes dados e quais não são. Outliers podem também representar, por exemplo, um comportamento não usual, uma tendência ou ainda transações fraudulentas (WITTEN, 2011).

2.2.3.3 Dados derivados

Em um conjunto de dados, as variáveis podem se relacionar umas com as outras. Sendo assim, se houver a necessidade de dados que não estejam disponíveis, é possível tentar obtê-los por meio de da transformação ou combinação com outras variáveis. Estes dados são chamados de dados derivados (WITTEN, 2011).

Um exemplo de um dado que pode ser calculado a partir de outro é a idade de um indivíduo. Neste caso, a idade pode ser calculada em função do dia atual e da data de nascimento (WITTEN, 2011).

2.2.4 Transformação dos dados

Transformação dos dados é a fase do KDD que antecede a fase de Mineração de Dados. Após serem selecionados, limpos e pré-processados os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos de aprendizado possam ser aplicados (WITTEN, 2011).

Em grandes corporações é comum encontrar computadores executando diferentes sistemas operacionais e diferentes Sistemas Gerenciadores de Bancos de Dados (SGDB). Estes dados que estão dispersos podem ser agrupados em um repositório único, de modo a facilitar a leitura e utilização deles (WITTEN, 2011).

Nesta etapa procura-se encontrar uma maneira de representar os dados. Estes dados foram agrupados nas etapas anteriores para um padrão que pode ser representado pela formatação ARFF *(Attribute-Relation File Format*) e posteriormente pode ser aplicado no WEKA para mineração dos dados (WITTEN, 2011).

A transformação de data também trata seis diferentes formas de tratar a data de entrada para mineração de dados. As formas, descritas nas seções a seguir, são: seleção de atributos, desratização de atributos, amostragem, transformação de classes para binário.

2.2.4.1 Seleção de atributos

Pode ser considerado o primeiro passo na transformação. Em algumas situações há muitos atributos a serem selecionados na base de dados, assim como

podem existir alguns os quais podem ser irrelevantes ou redundantes. Assim pode-se precisar que estes atributos sejam processados para selecionar um grupo menor garantindo assim uma melhor performance (WITTEN, 2011).

2.2.4.2 Discretização

A discretização deve ocorrer para que os atributos numéricos de grande escala possam ser divididos em valores menores com uma amplitude distinta para o uso posterior na mineração de dados (WITTEN, 2011).

No entanto, quando um atributo discretizado é derivado de um atributo numérico, seus valores são ordenados e o tratamento como nominal descarta essas informações, cada atributo discretizado é declarado como sendo do tipo ordenado (WITTEN, 2011).

2.2.4.3 Amostragem

A amostragem envolve grande volume de dados, sendo assim é necessário selecionar uma pequena amostra para ser processada. Esta deve ser uma instancia da base de dados original (WITTEN, 2011).

2.2.4.4 Transformação

É uma prática comum transformar problemas multiclasse em múltiplos de duas classes (WITTEN, 2011). Nesta etapa, o conjunto de dados é decomposto em problemas de duas classes, o algoritmo é executado em cada dado e as saídas dos classificadores resultantes são combinadas de modo a produzir diversos conjuntos de dados de classes discriminando cada classe contra a união de todas as outras.

Essa técnica é comumente chamada de um contra os outros (*one-vs-rest*, e erroneamente de *one-vs-all,* um contra todos). Para cada classe, um conjunto de dados é gerado contendo uma cópia de cada instância nos dados originais, mas com um valor de classe modificado. Se as instâncias corresponderem a classe associada ao conjunto de dados, elas serão marcadas como sim; caso contrário não (WITTEN, 2011).

2.2.5 Mineração de Dados

Todas as etapas do processo de KDD possuem grau elevado de importância para o seu sucesso. Entretanto, a etapa de Mineração de Dados (MD) recebe o maior destaque na literatura (WITTEN, 2011).

A MD tem como objetivo construir hipóteses, modelos, grupos etc. Considerando o paradigma simbólico: regras de produção e árvores de decisão. Para isso, existem diversos algoritmos e sistemas de aprendizado, tais como C4.5, Regras de Associação, entre outros.

A técnica de MD também procura descobrir e descrever padrões de estruturas de dados, como uma ferramenta para ajudar a explicar os dados e conseguir prever resultados a partir dela (WITTEN, 2011).

Segundo Oliveira (2009), a MD é reconhecida pela execução de suas diversas tarefas, sendo as mais comuns:

- Descrição: descreve padrões e tendências reveladas pelos dados, geralmente oferecendo uma interpretação dos dados obtidos;
- Classificação: busca determinar a qual classe que um registro pertence;
- Estimação ou Regressão: pode se estimar o valor de uma variável analisando as demais;
- Agrupamento: aproxima os registros similares, identificando assim seus Grupos;
- Associação: identifica quais atributos estão relacionados.

Uma vantagem da MD sobre outros tipos de análise de dados é a necessidade de o conjunto de dados ser analisado em um só local, tornando mais precisa a descoberta informações de forma automática, visto que os dados que estão sendo guardados já são automaticamente preparados para a mineração.

2.2.5.1 Método de classificação

A classificação é uma tarefa popular na descoberta de conhecimento em bases de dados. Ela tem como objetivo encontrar uma função que mapeia um conjunto de registros em um conjunto de classes pré-definidas. Uma vez obtida esta função, ela

pode ser aplicada a novos registros para prever a classe em que estes se enquadrariam ou se enquadram (QUINLAN, 1996).

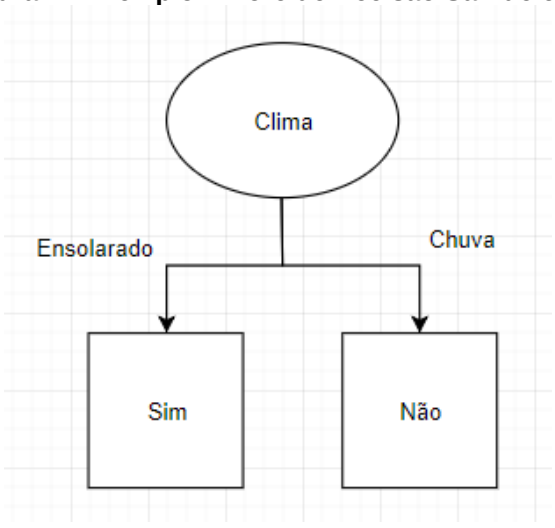
É o processo de produzir regras que possam ser executadas sem ordem de relevância para definir uma classe, um conjunto de rótulos (WITTEN, 2011). Em outras palavras o método de classificação equivale à obtenção de regras baseando em observações de estados, se e então. Como por exemplo para um animal ser uma ave ele deve possuir um corpo coberto de pelos e possui um bico, podendo assim gerar a regra, se corpo coberto e possui bico então é uma ave.

2.2.5.2 Algoritmo C4.5

O C4.5 é um algoritmo de divisão e conquista que gera uma árvore de decisão a partir de um conjunto de regras, classificadores neste formato podem ser lidos como regras do tipo *se então*, os quais possuem o objetivo de representar o conhecimento de um dado grupo (MITCHELL,1997).

A árvore de decisão é a junção de um conjunto de regras para determinar um grupo de classificadores. Esta é formada por um nó raiz, o qual indica uma regra base para a árvore, ou seja, o atributo que melhor separa por si só os exemplos a serem classificados e vários nós folhas os quais resultam de cada regra (MITCHELL, 1997). A Figura 2 exemplifica a decisão baseada no clima, onde caso esteja ensolarado a reposta sim, caso contrário não.

Figura 2 - Exemplo Árvore de Decisão Sair de casa



Fonte: Adaptado de (WITTEN, 2011)

Árvores de decisão são caracterizadas por serem robustas, resistentes a ruídos nos dados, de maneira a serem utilizadas para classificar valores altos e baixos assim como valores contínuos para determinado atributo (QUINLAN, 1996).

2.2.5.3 Pseudocódigo do Algoritmo C4.5

O processo de criação da árvore de decisão é a partir de um conjunto de dados de treinamento *T*. Com T possuindo j possíveis valores para as classes sendo $A = \{a_1, ..., a_i\}$ o conjunto contendo i atributos, cujos exemplos são rotulados com as m classes $C = \{c_1, ..., c_m\}$. É feito da seguinte forma (QUINLAN, 1996):

1. Se *T* contém um ou mais exemplos, todos pertencentes à mesma classe c_x, a árvore de decisão irá gerar uma folha pela classe c_j;

2. Caso o conjunto de treinamento T for vazio, a árvore de decisão será formada por apenas um nó folha cuja classe associada a ele será determinada pela informação mais frequente existente no conjunto C;

3. Se T contiver exemplos que pertençam a diferentes classes c, deve-se escolher um atributo ai pertencente ao conjunto de atributos A, deve-se particionar T em subconjuntos $\{T_1, ..., T_j\}$, onde T_i contém todos os exemplos que possuem os *k* valores possíveis para o atributo ai, executando novamente o processo e voltando ao item 1 para cada para cada exemplo pertencente ao conjunto *T*.

Este processo será repetido até que todos os exemplos possam ser classificados pela árvore. Se uma sub-árvore tenha uma classificação inferior a um nó folha, está será removida, sendo assim atribuído a este nó a característica mais comum associada aos exemplos de treinamento, isso é caracterizado como processo de *poda* e consequentemente reduzindo o tamanho da árvore (QUINLAN, 1996). Com a presença de valores contínuos, um limiar *t* que produz maior ganho de informação é determinado de forma automática pelo algoritmo C4.5. Os valores devem estar ordenados e terem as classes resultantes associadas a eles, assim calculando o ganho de informação para o primeiro elemento em relação aos demais, em seguida, é calculado o ganho para um conjunto formado pelo primeiro e o segundo em relação aos demais e assim por diante.

Para separar os subconjuntos de *T*, o principal critério é o de ganho de informação. Tal consiste em medir a frequência *freq (C, s)* na qual um conjunto

qualquer de exemplos (*s)* pertencentes a mesma classe (*C*), sendo *S* o número total de exemplos, conforme é mostrado na Equação 1.

$$GanhoInformação = \frac{Fr(Cj, s)}{S} \quad (1)$$

2.2.5.4 Exemplo de Aplicação C4.5

Tendo como exemplo o ato de jogar ou não bola fora de casa é necessário que tenhamos alguns atributos para a decisão, estes são tempo, temperatura, umidade, vento assim podendo decidir se podemos ou não jogar bola fora de casa. Para a criação de um conjunto de treinamento é preciso que cada exemplo possua todos os atributos respondidos e por fim deve saber se pode ou não jogar. Assim possibilitando a criação da Tabela 1.

Tabela 1 - Conjunto de Treinamento

Exemplo	Tempo	Temperatura	Umidade	Vento	Classe
1	Ensolarado	Média	Normal	Sim	Joga
2	Ensolarado	Alta	Alta	Sim	Não Joga
3	Ensolarado	Alta	Alta	Não	Não Joga
4	Ensolarado	Média	Alta	Não	Não Joga
5	Ensolarado	Baixa	Normal	Não	Joga
6	Nublado	Alta	Alta	Não	Joga
7	Nublado	Alta	Normal	Não	Joga
8	Nublado	Baixa	Normal	Sim	Joga
9	Nublado	Média	Alta	Sim	Joga
10	Chovendo	Baixa	Normal	Sim	Não Joga
11	Chovendo	Média	Alta	Sim	Não Joga
12	Chovendo	Baixa	Normal	Não	Joga
13	Chovendo	Média	Alta	Não	Joga
14	Chovendo	Média	Normal	Não	Joga

Fonte: Adaptada de (WITTEN, 2011)

Com a Tabela 1 é possível determinar que a frequência de exemplos positivos são 9/14 ou 0.642 e 5/14 ou 0.357 de exemplos negativos. Deve-se criar uma árvore

de decisão baseando-se nos atributos dados pela tabela, onde o classificador é jogar ou não jogar.

A busca pela informação deseja de um atributo, a qual pertence uma classe, é calculada pela equação *info(S)*, que busca a frequência de exemplos que resultam em casos positivos presentes no conjunto de treinamento e multiplica-se pelo logaritmo na base 2 dos mesmos exemplos positivos, somando com exemplos negativos, que são obtidos da mesma forma, conforme pode ser visto na Equação 2.

$$Info\ (S) = -\sum_{j=1}^{k} \frac{freq(\ C,s)}{S}\ x \log_2(\frac{freq(C,s)}{S})bits \quad (2)$$

A entropia(E), de um conjunto S é calculada aplicando a equação acima medindo assim a quantidade de informação necessária para que seja possível identificar as classes de um determinado caso num conjunto de treinamento, calculando assim a informação do conjunto de treinamento com a função *info*(t) (MITCHELL, 1997).

O critério de ganho, pode ser feito utilizando a seleção de um teste que maximize o ganho de informação, este é medido pela entropia menos a informação de T do atributo x que pode ser visto na Equação 3.

$$Ganho(X) = Entropia(E) - Info_x(T) \quad (3)$$

Sendo assim a entropia dos dados incide de:

$$Info(T) = -\frac{9}{14} * log_2\left(\frac{9}{14}\right) - \frac{5}{14} * log_2\left(\frac{5}{14}\right) = 0.940\ bits \quad (4)$$

Desta forma é possível calcular o ganho para cada atributo da base de dados, escolhendo assim o atributo que melhor separa por si só as informações.

Após calcular as informações de cada atributo é possível calcular o ganho para eles:

$$\text{Ganho (Tempo)} = 0{,}940 - 0{,}694 = 0{,}246 \quad (5)$$

$$\text{Ganho (Temperatura)} = 0{,}940 - 0{,}890 = 0{,}05 \quad (6)$$

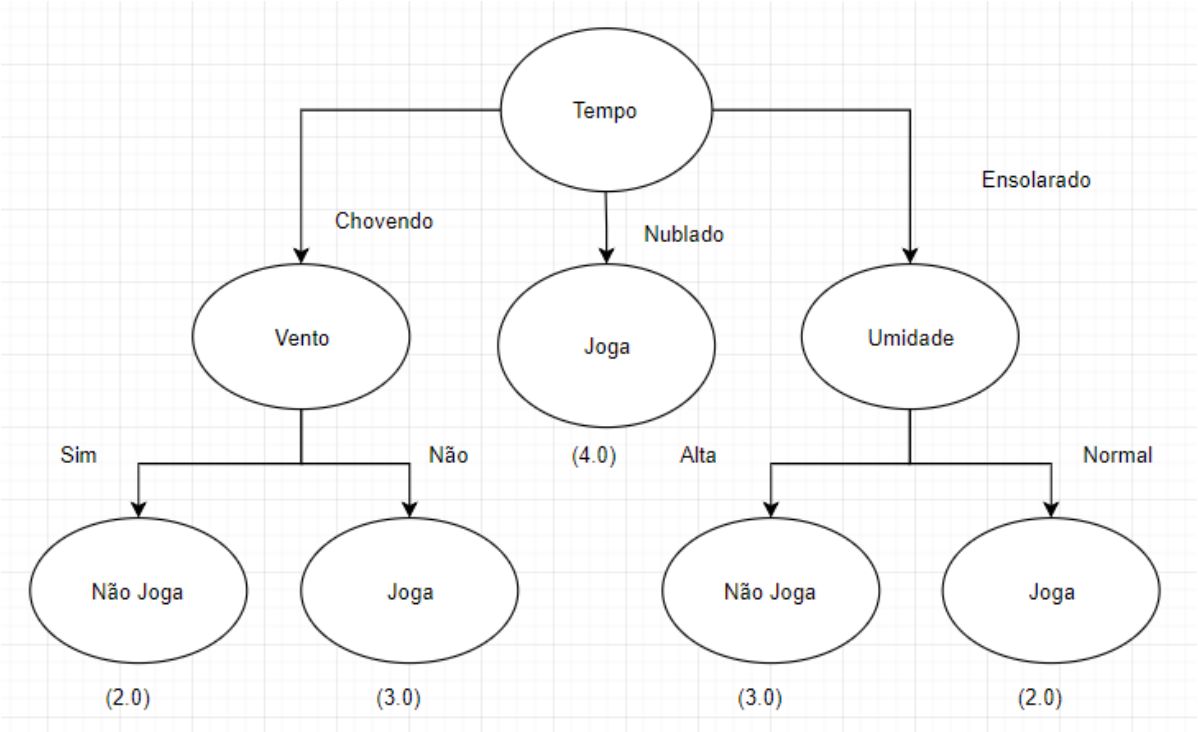
$$\text{Ganho (Umidade)} = 0{,}940 - 0{,}788 = 0{,}152 \quad (7)$$

$$\text{Ganho (Vento)} = 0{,}940 - 0{,}892 = 0{,}048 \quad (8)$$

O valor mais alto de ganho (X) é escolhido como raiz da arvore, neste exemplo o atributo tempo.

Ao executar o algoritmo C4.5 é gerada uma estrutura que pode ser vista na Figura 3. Cada valor gerado entre parênteses corresponde ao número de exemplos bem classificados / mal classificados.

Figura 3 - Exemplo da Árvore de Decisão para Jogar



Fonte: Adaptada de (WITTEN, 2011)

A Figura 3 mostra que está árvore gerada é uma versão simplificada dos atributos da fase de teste pois nesta árvore o atributo temperatura não influência na classificação.

A mesma árvore pode ser escrita em formato de regras, sendo classificada ao lado e a precisão da regra em porcentagens, como por exemplo a Regra 1, pode ser liga: *se o tempo estiver nublado,* então tem a identificação que pode jogar.

O resultado do C4.5 é mostrado também por uma matriz de confusão. Nela são mostrados como o exemplo foi classificado e sua classificação correta.

Tabela 2 - Exemplo Matriz de Confusão

A	**B**	
9	**0**	**A = Joga**
0	**5**	**B = Não Joga**

Fonte: Autoria Própria

No exemplo, 8 amostras do tipo ‘A’ foram classificadas corretamente como ‘A’ e as 5 amostras do tipo ‘B’ foram classificadas como ‘B’, onde demonstra que o classificador não obteve erro na hora de fazer a classificação da amostra.

2.2.5.5 Método de *Bagging* com o algoritmo C4.5

Bagging é outra forma de obter uma árvore de decisão sugerida por (BREIMAN,1996). O método baseia-se na geração e combinação de múltiplos classificadores que são obtidos a partir de diferentes amostras de um conjunto de treinamento (SILLA, 2005).

O método *Bagging* é eficiente quando é feita a combinação de classificações de modelos diferentes, explorando desta forma a instabilidade do método de aprendizagem, fazendo com que um modelo complemente o outro (WITTEN, 2011). Para o método *Bagging* o peso é sempre o mesmo para diferentes modelos.

Os classificadores serão treinados de forma independente por diferentes conjuntos de treinamento por meio do método de inicialização. Para construí-los é necessário montar *k* conjuntos de treinamento idênticos e replicar esses dados de treinamento de forma aleatória e construir *k* instâncias independentes por amostragem. Fazer múltiplas vezes sem que perca os dados de inserção. Em seguida, deve-se agregar as *k* instancias por meio de um método de combinação apropriada, tal como a maioria de votos (BREIMAN, 1996).

2.2.5.6 Método de *Boosting*

O *Boosting* de forma semelhante ao *Bagging* tenta melhorar a precisão dos algoritmos de aprendizagem, reduzindo a taxa de erro por meio de técnica que busca combinar classificadores, os quais são geradores em um processo que busca gerar classificadores complementares.

Este usa conjuntos para gerar arvores de conjuntos treinados de dados, a principal diferença entre os dois métodos ocorre na performance dos dados uma vez que o Boosting melhora os dados de baixa performance e o *Bagging* faz reposição com os mesmos valores tentando criar conjuntos diferentes (LANTZ, 2013).

AdaBoost ou *adaptive boosting* corresponde a um algoritmo de *boosting* proposto em 1997. Este algoritmo é baseado na ideia de gerar maiores números de treinamentos com amostragem de classificadores fracos e aumentar a performance destes. Boosting e Bagging são comumente usados para fazer divisão e conquista usando árvores de decisão e regras, pois conseguem ser aplicados generalizadamente em todos os algoritmos de *machine learning* (LANTZ, 2013).

2.3 FERRAMENTA WEKA

O WEKA tem em sua aplicação as ferramentas necessárias para filtragem da amostra com discretização, possui também repositório com diversos tipos de algoritmos para mineração de dados como classificadores, geradores de árvores etc.

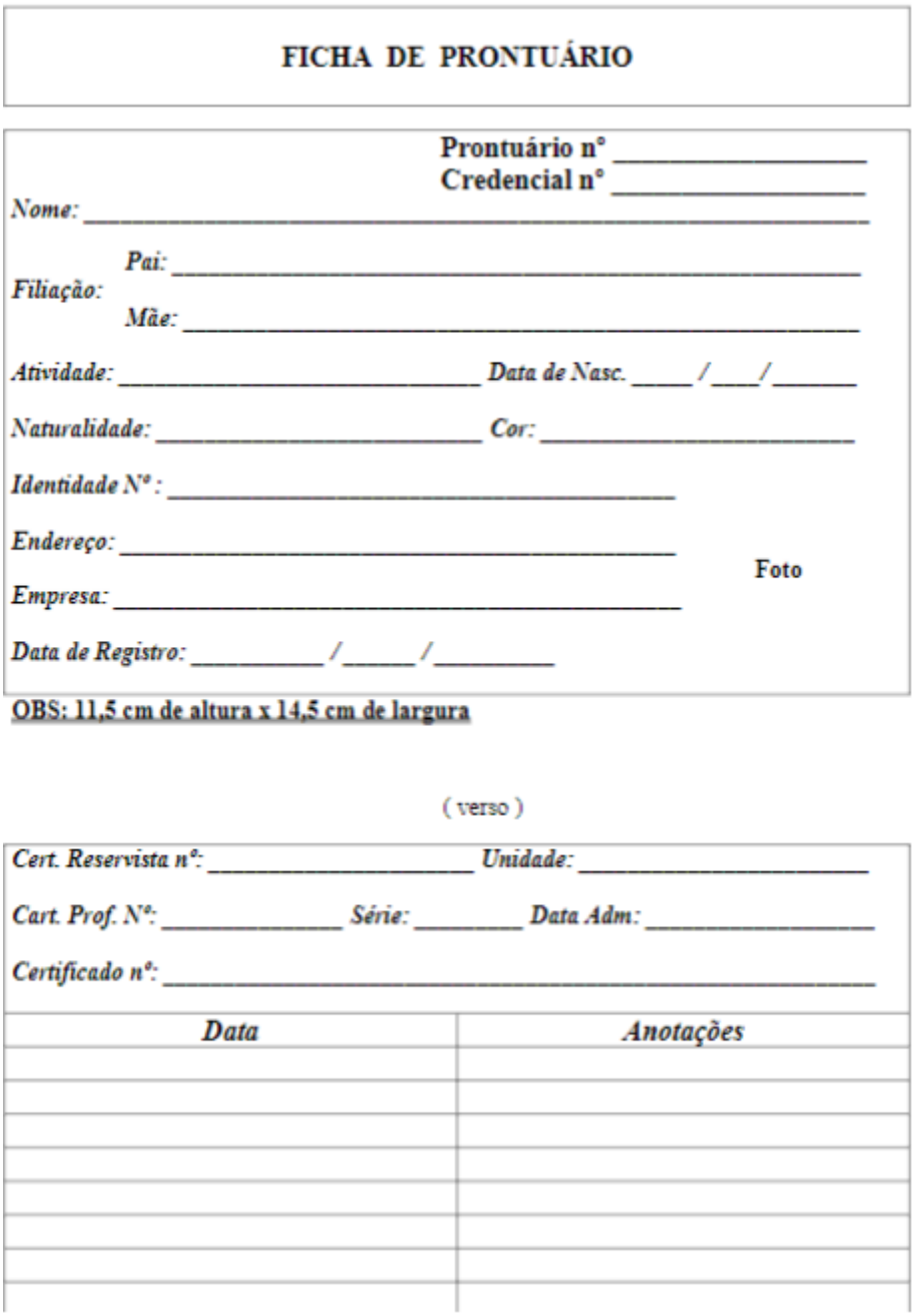
Para este trabalho usa-se a ferramenta para a discretização dos dados e para a mineração destes com o objetivo em analisar a resposta da mineração e tentar descobrir um novo conhecimento.

2.4 PRONTUÁRIO ELETRÔNICO

A palavra prontuário origina-se do latim *promptuarium* e significa “Lugar onde são guardadas coisas de que se pode precisar a qualquer momento” (THOMAZ, 2018). Na medicina, o prontuário é uma documentação legal permanente das informações relevantes para o gerenciamento do cuidado de saúde de um cliente (CARVALHO, 2012).

Os prontuários podem ser eletrônicos ou em papéis. Os prontuários eletrônicos possuem como vantagens: a agilidade no preenchimento dos documentos; segurança dos dados; atualização em tempo real e a portabilidade. Já os prontuários feitos em papel, podem ser visualizados na **Erro! Fonte de referência não encontrada.**.

Figura 4 - Prontuário Físico



FICHA DE PRONTUÁRIO

Prontuário nº _____
Credencial nº _____

Nome: _____

Filiação: *Pai:* _____
Mãe: _____

Atividade: _____ *Data de Nasc.* ____ / ___ / _____

Naturalidade: _____ *Cor:* _____

Identidade Nº: _____

Endereço: _____

Foto

Empresa: _____

Data de Registro: _____ / _____ / _____

OBS: 11,5 cm de altura x 14,5 cm de largura

(verso)

Cert. Reservista nº: _____ *Unidade:* _____

Cart. Prof. Nº: _____ *Série:* _____ *Data Adm:* _____

Certificado nº: _____

Data	*Anotações*

Fonte: (THOMAZ, 2018).

Têm como vantagem sua validade jurídica, uma vez que o papel apresenta valor documental, e o baixo nível de investimento para mantê-los (ALVES, 2014).

O registro de informação de pacientes por muito tempo foi feito em papel. Contudo, esta forma de armazenamento de dados está propensa a problemas como extravio, corrosão do papel por efeito do tempo, quebra de sigilo, falta de mobilidade de dados e entre outros (SILVA, 2007). Problemas os quais podem até mesmo comprometer o atendimento e o diagnóstico seguro. Em razão dos avanços na

medicina e na tecnologia na área da saúde foi possibilitado ao longo do tempo o desenvolvimento do prontuário eletrônico do paciente, o qual deve ser capaz gravar, guardar e disponibilizar os dados sobre o paciente e seu tratamento (SILVA, 2007).

Os prontuários eletrônicos começaram a ser desenvolvidos nos anos 60 nos Estados Unidos, estes se restringiam a grandes hospitais em parcerias com universidades. No ano de 1991 o *Institute of Medicine* (IOM) publicou um relatório (RICHARD, 1991) pedindo a eliminação de registros de pacientes em papel dentro de um período de 10 anos (THOMAS, 2018). Thomaz também cita que no Brasil, a regulamentação do Prontuário Eletrônico foi implementada em 2002, quando o Conselho Federal de Medicina (CFM) definiu suas características gerais na resolução 1638.

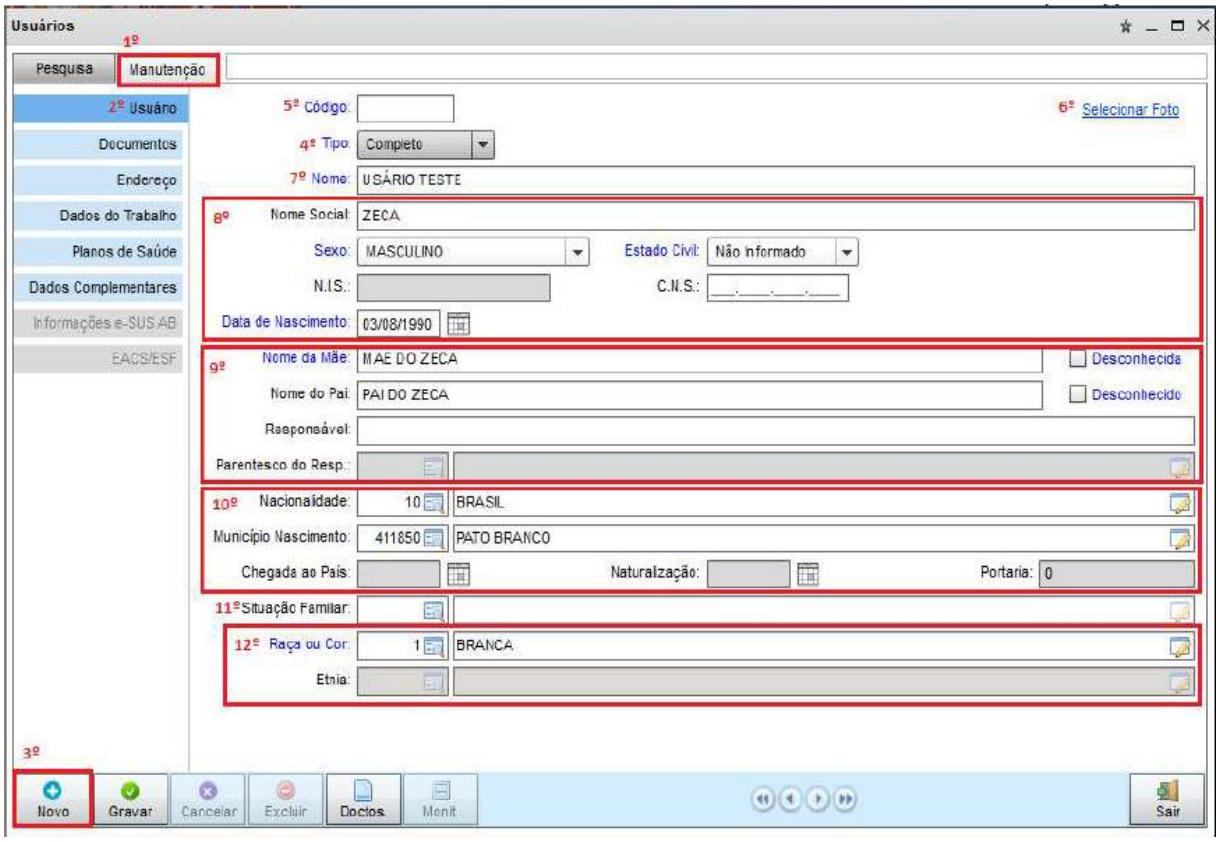
A inserção de prontuários eletrônicos no Sistema Único de Saúde (SUS) foi determinada na portaria do Ministério da Saúde (2011), a qual prevê que todas as Unidades Básicas de Saúde (UBS) utilizem os prontuários eletrônicos do paciente, sob a pena de corte de repasses (ALVES, 2018). A medida tem como objetivo criar um cadastro de pacientes o qual poderá ter acessado garantindo e que os profissionais tenham acesso aos dados, seja em consultas ou intervenções possibilitando um melhor atendimento do paciente.

Para o acesso à informação se faz necessário a certificação digital, garantindo assim que apenas pessoas autorizadas serão capazes se visualizar informações sigilosas sobre o paciente.

2.4.1 Criação do Prontuário Eletrônico

Para criação do prontuário eletrônico do paciente é necessário ser realizado previamente um cadastro completo deste para sua identificação, dados como o nome, CPF, sexo, cor, data de nascimento, estado civil, endereço são requisitados. Na Figura 5 é possível visualizar a tela de criação de prontuário eletrônico do paciente no IDS-Saúde, software desenvolvido pela empresa IDS.

Figura 5 - Exemplo de um Cadastro de Paciente em um Prontuário Eletrônico

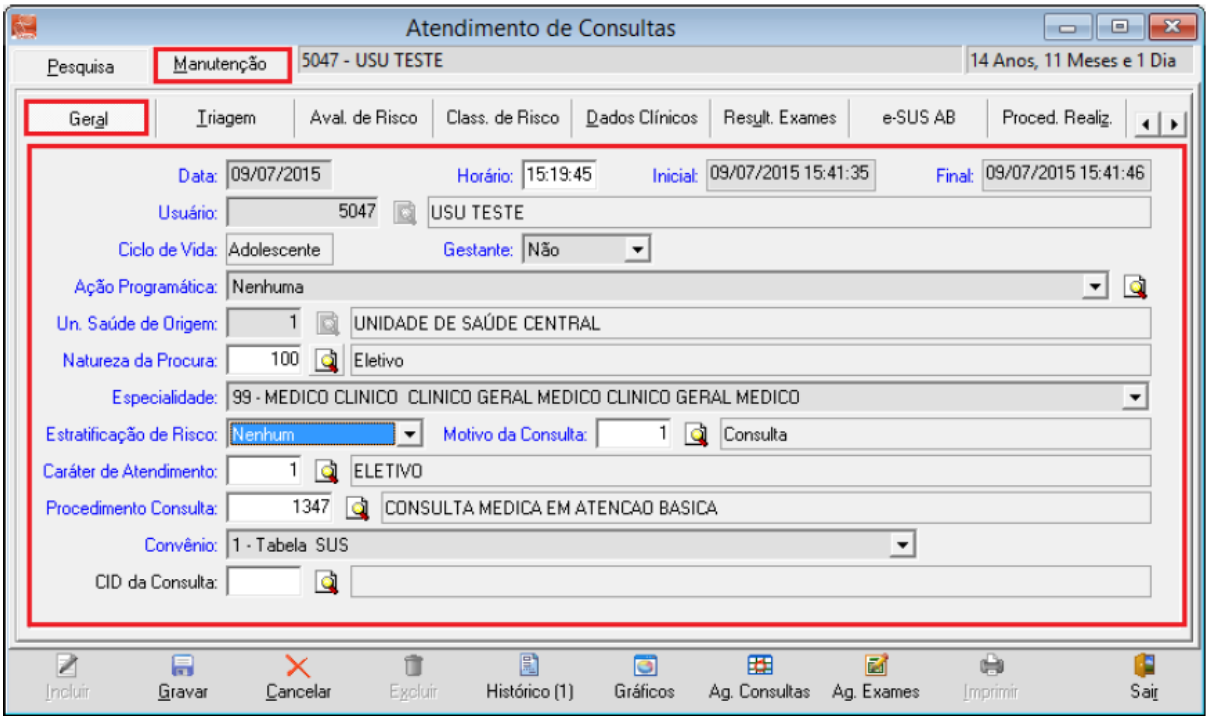


Fonte: (IDS, 2015)

Uma vez cadastrado o paciente, o seu prontuário eletrônico estará pronto para uso. Já durante o atendimento de consultas, cabe ao usuário do sistema, médico(a) ou enfermeiro(a), incluir informações quanto a anamnese, triagem, exames, hipóteses diagnosticas, diagnósticos definitivos, tratamentos realizados e prescrição de medicamentos (IDS, 2015).

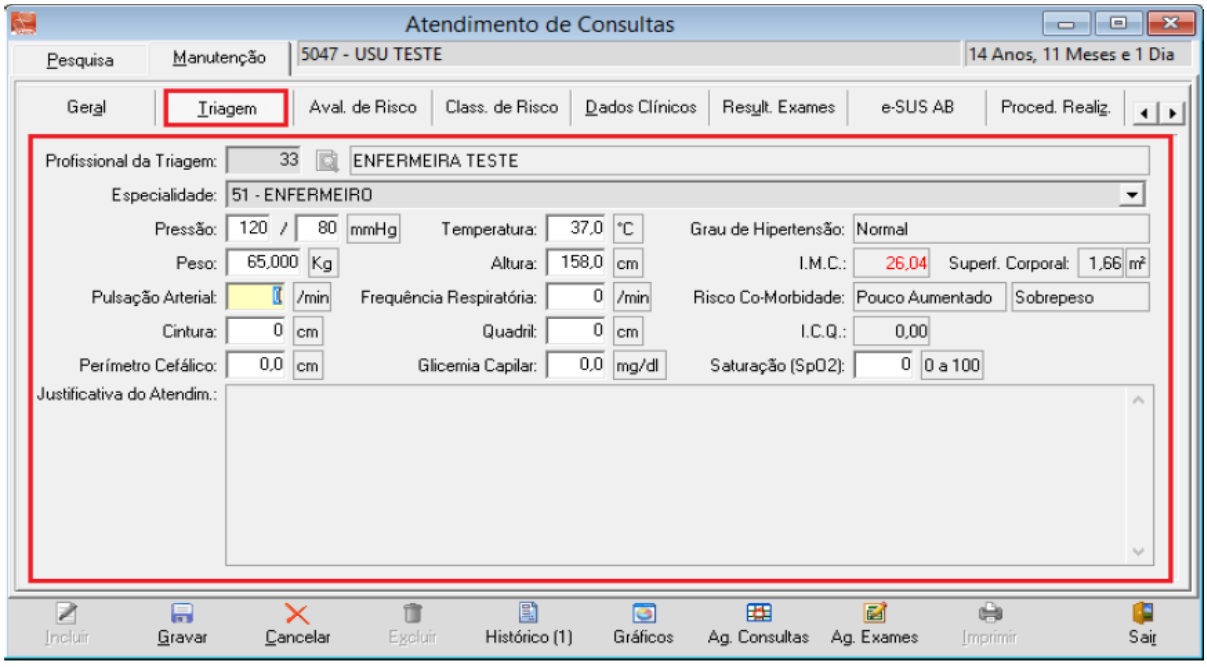
Além destas informações podem ser inseridos outros dados conformes as necessidades do usuário, como alergias do paciente, anotações da consulta etc. O modo em que estes dados são inseridos no sistema ao se realizar o atendimento do paciente pode ser visto na Figura 6 e Figura 7.

Figura 6 - Exemplo de Dados Gerais de um prontuário



Fonte: (IDS, 2015)

Figura 7 - Exemplo de dados de triagem de um prontuário



Fonte: (IDS, 2015)

Atualmente existem diversos sistemas de prontuários eletrônicos, estes também buscam atender as mais diversas funcionalidades além do obrigatório, cabendo aos clientes escolher o sistema que melhor atenda suas expectativas e

necessidades, possibilitando a melhoria na eficiência em atendimento, redução de erros de registro e até mesmo aumento de produtividade.

2.4.2 Softwares para Prontuários Eletrônicos

Os softwares mais conhecidos hoje são o *iClinic* e o *Prontmed* para criação e cuidados de prontuários eletrônicos para clínicas, estes dão aos clientes um jeito mais fácil de controlar a sua clínica para poder aumentar a produtividade. Estes softwares por estarem dentro de uma mesma clientela acabam possuindo os mesmos produtos base indicados na Tabela 3.

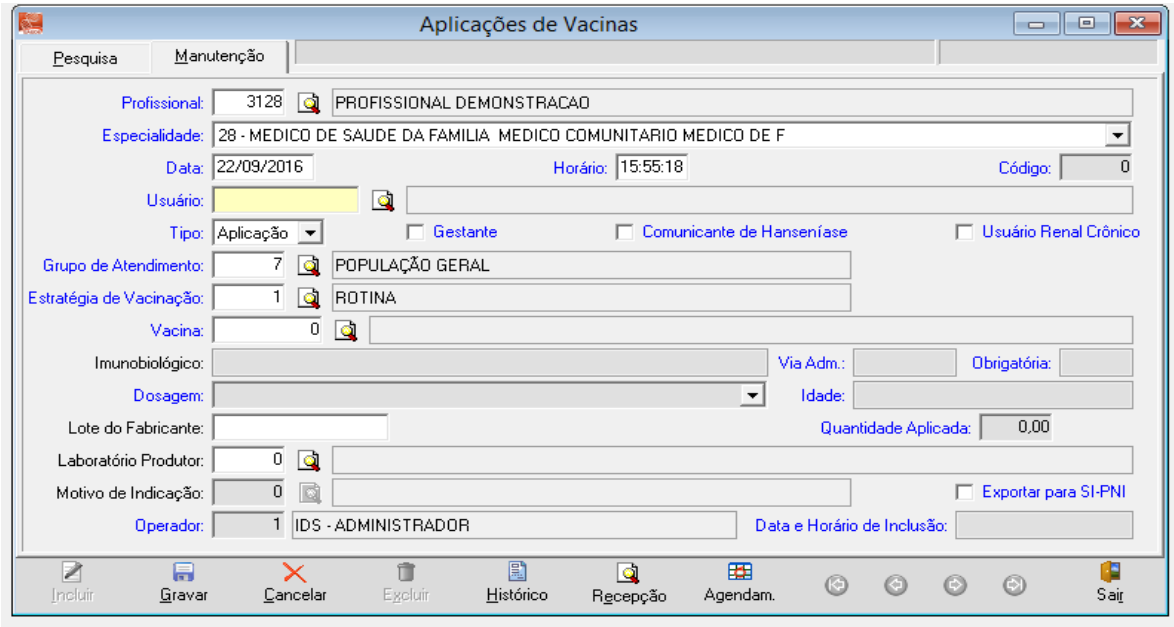
Tabela 3 - Comparativo de sistemas de Prontuário Eletrônico

Prontuário Eletrônico	iClinic	ProntMed	IDS
Nome	X	X	X
Família	-	-	X
Convênio	X	X	-
Estado Civil	X	X	X
Hábitos	X	X	X
Controle Financeiro	X	X	-
Secretaria Eletrônica	X	X	-
Marketing Médico	X	-	-

Fonte: Autoria Própria

Um sistema de saúde deve possuir vários tipos de telas para que possam conter um prontuário bem explicado. Estas telas têm o objetivo de separar os atendimentos dentro do SUS onde, as telas podem ser para aplicações de vacinações, triagens e o prontuário da família do paciente. Um exemplo de tela específica de vacinas pode ser visto na Figura 8.

Figura 8 - Dados referentes à aplicação de vacinas



Fonte: (IDS, 2015)

A folha de prescrição médica Anamnese e exame físico são os dados do exame inicial realizado pelo médico, incluindo história familiar, diagnósticos confirmados e plano de cuidados.

O registro de medicamentos é a documentação exata de todos os medicamentos administrados como data, hora, dose, via de administração, assinatura de quem preparou e administrou.

O sumário de alta contém a condição do cliente, evolução, prognóstico, reabilitação e necessidades de ensino no momento da alta do hospital.

2.5 TRABALHOS RELACIONADOS

Está sessão mostra alguns trabalhos já feitos na área de descoberta de conhecimento de bases de dados com a área da saúde. Para isto foi feito uma pesquisa nas bases de dados como IEEE, Periódicos Capes. Para a pesquisa foi utilizado as palavras-chaves WEKA, Mineração de Dados, KDD, prontuário eletrônico e área da saúde.

Foi realizada também uma busca também no Google Scholar, por este ser um local de busca de artigos de diversas fontes. Durante a pesquisa foram escolhidos

trabalhos feitos posteriores a data de 2010 para que estes apresentem um conteúdo mais atualizado da área de computação.

Vilarinho R. (2017) utiliza algoritmos de mineração para obtenção de informações úteis relativas a casos de Dengue nos munícipios brasileiros. Este trabalho utiliza bases de dados providas do Governo Federal, IBGE e Atlas. Para este trabalho Renato utiliza os algoritmos C4.5 e K-*means*, que foram implementados utilizando a ferramenta WEKA. Este trabalho teve o objetivo de encontrar métricas para identificação de grupos denominados “Muito Baixo”, “Baixo”, “Médio”, “Alto” e “Muito Alto”. (Vilarinho, 2017).

Trindade et al. (2012) aplicou o KDD para a identificação de padrões de comportamento das Hepatites Virais nas bases de dados do SINAN (Sistema de Informação de Agravos e Notificações) do Sistema Único de Saúde - Governo Federal do Brasil, objetivando subsidiar ações de controle e prevenção da doença. Ele também utiliza o algoritmo C4.5 e faz uso de recursos visuais, como gráficos e mapas, para facilitar o entendimento dos padrões encontrados para o especialista. Em seu trabalho é destacado que apenas 65 atributos dos 134 selecionados puderam ser utilizados devido à ausência do dado, sugerindo falhas durante a coleta dos dados.

Carvalho R. D.; Escobar L. F. A.; Tsunoda D. (2014) utiliza da literatura para mostrar pontos de atenção no uso de mineração de dados na área da saúde detalhando autores importantes, métodos utilizados. Também mostra que a predominância de publicações encontradas que relatam adoção da mineração de dados para área da clínica, evidencia um espaço importante da utilização desta técnica na rotina para potencializar a eficiência dos tratamentos.

Em seu trabalho (CARVALHO et al, 2014) também faz uso de recursos visuais, como tabelas e gráficos, para evidenciar os principais pontos de atenção que foram vistos nos trabalhos revisados, os principais pontos são a previsão da ocorrência de eventos numa determinada população e o auxílio ao planejamento em saúde.

Feuser R. (2017) aplica os processos do KDD para prontuário eletrônico de paciente oriundo, em seu trabalho ele usa regras de associação e mineração de dados, utilizando a ferramenta WEKA para mineração e visualização dos resultados. Em seu trabalho fica evidenciado que o grupo CID J20 (Outras infecções agudas das vias aéreas inferiores) possui uma confiança de 92% quando associados com o grupo R50 (sinais como febre de origem desconhecidas). A alta taxa de detecção destes CID’s deve evidenciar aparentemente que estes casos possuem maiores ocorrências.

3 DESENVOLVIMENTO

3.1 BASE DE DADOS

A base de dados utilizada neste trabalho foi disponibilizada através de um projeto com a participação da Secretária Municipal de Saúde de Pato Branco, a Secretaria de Ciência e Tecnologia, a universidade Tecnológica Federal do Paraná, Campus Pato Branco, e a empresa desenvolvedora do sistema de prontuário (IDS Desenvolvimento de Software e Assessoria Ltda.). Esta foi primeiramente disponibilizada para Rodrigo Feuser para seu trabalho de conclusão de curso no ano de 2017, este trabalho usa a mesma base de dados, e algumas funções também feitas por Rodrigo Feuser.

Dados pessoas, como nome, RG, CPF, telefone e outros não fazem parte deste projeto, isso garante que em momento algum, as pessoas envolvidas no projeto tiveram acesso a identificação dos pacientes e de seus responsáveis.

A base de dados foi entregue em formato de backup para banco de dados PostgreSQL.

3.1.1 Recuperação da base

Para a recuperação do arquivo é necessário possuir instalado a ferramenta PostgreSQL, para este projeto foi usado a versão PostgreSQL 11. Assim que instalado é necessário que seja feita a conexão ao banco usando o terminal de comando e então escolher o banco a ser recuperado, conforme indicado na Figura 9.

Figura 9 - Recuperar Banco

```
1 \i C:/Users/Kaique/arquivo.sql
```

Fonte: Autoria Própria

Com o uso do \i é realizada uma verificação do arquivo no local informado no computador para saber se este é um arquivo o qual pode ser recuperado no banco de dados, com a execução desse comando deve ser recuperado toda a base de dados adicionando um esquema ao banco público e criando um banco chamado Winsaude.

3.1.2 Base

A divisão de tabelas entre os dois esquemas público e Winsaude são: público possui trinta e quatro tabelas incluindo três tabelas com dicionário do banco de dados, e o Winsaude possui cinco tabelas.

3.2 PROCESSO KDD

Para que seja possível a aplicação do KDD precisa-se criar uma cópia da tabela com que iremos usar para fazer o processamento, neste trabalho é utilizada a tabela SAFICATE, ou seja, todo o processamento dos dados será focado nos prontuários eletrônicos do paciente e para a etapa de transformação serão usadas as demais tabelas para que seja possível resgatar todos os dados necessários.

3.2.1 Seleção dos Dados

A listagem de tabelas que serão utilizadas para este projeto pode ser vista com o comando SQL de seleção. Por meio deste comando é gerado Tabela 4 ilustrada a seguir.

Tabela 4 - Retorno consulta SQL

	tabcodigo smallint	tabnomfis character varying (9)	tabdescri character varying (50)	tabchprim character varying (80)	tabsigla character varying (3)	tabforcad smallint
1	108	SASEXOS	Sexos	SEXCODIGO	SEX	0
2	107	SAUSUARI	Usuários	USUCODIGO	USU	1
3	103	SABAIDIS	Bairros e Distritos	BDICODIGO	BDI	1
4	157	SAFICATE	Fichas de Atendimentos	USUCODIGO;FATCODIGO	FAT	0
5	674	SASUSUSU	Informações de e-SUS dos U...	USUCODIGO	ESU	0

Fonte: Autoria Própria

A tabela 4 contém os seguintes campos:

- SASEXOS: Informa três tipos de situações para sexo dos usuários, masculino, feminino e indiferente;
- SAUSUARI: Tabela de usuários, possui os campos relativos aos pacientes, ou seja, usuários do sistema de saúde SUS;

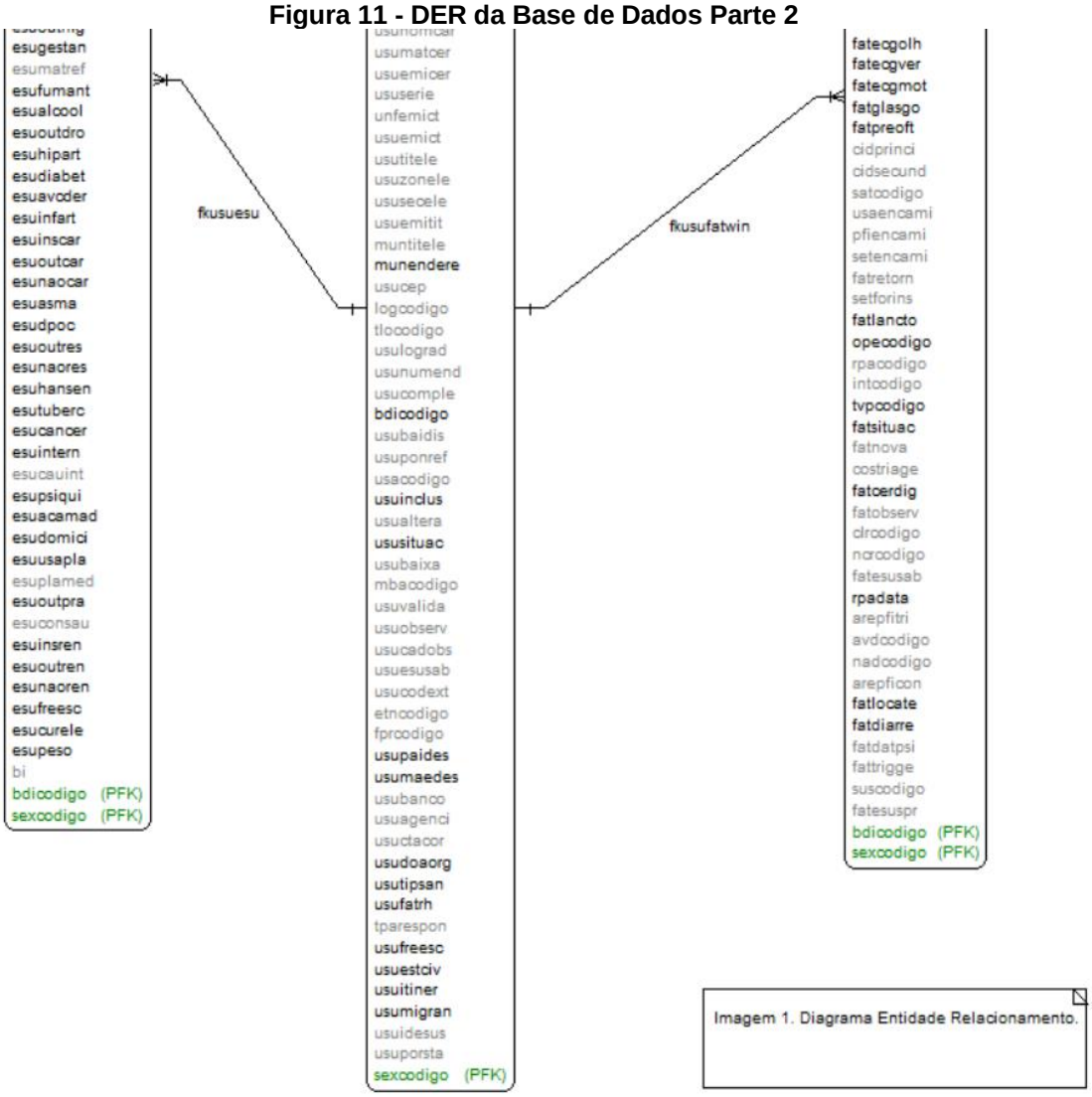
- SABAIDIS: Traz as informações referentes aos bairros e localidades dos usuários do sistema de saúde. Os logradouros, ruas e numerações não são usados para não identificar local de moradia dos pacientes;
- SAFICATE: Esta tabela possui o prontuário eletrônico do paciente. Nela estão todos os detalhes do atendimento, desde a triagem, o atendimento médico e a alta ou encaminhamento para casa hospitalar;
- SASUSUSU: As informações de e-SUS, que é um questionário de situação socioeconômica do paciente Além de históricos de doenças e hábitos de vida dos usuários.

O Diagrama de Entidade e Relacionamento (DER) completo pode ser visto na Figura 10 e Figura 11.

Figura 10 - DER da Base de Dados Parte 1

Fonte: (FEUSER, 2017)

Figura 11 - DER da Base de Dados Parte 2



Fonte: (FEUSER, 2017)

3.2.1.1 Seleção das Colunas

Para este trabalho é necessário a escolha de algumas colunas da tabela SAFICATE onde é possível criar um modo de fazer identificação de perfis com o enfoque em doenças. Isso faz com que seja obrigatório o uso da coluna cidprinci que possui os valores das doenças identificadas no prontuário eletrônico. Foi feita uma função para descobrir a faixa etária, o tipo de sexo de cada indivíduo e foram escolhidas algumas colunas relacionadas ao censo que é feito durante a criação do prontuário, tais como:

- Bairro;
- Altura;
- Peso;
- Frequência escolar;
- Frequenta Benzedeira;
- Possuir plano de saúde;
- Gestante;
- Hipertensão arterial;
- Diabetes;
- Asma;
- Alcoólatra;
- Fumante;
- Infarto; e
- AVC ou Derrame.

Estes atributos selecionados são possíveis posteriormente fazer a parte de transformação de dados.

3.2.1.2 Criação da Tabela

A criação de uma tabela é necessária para que possa ser feito os processos de KDD sobre esta. A tabela que deve ser criada será igual a tabela SAFICATE e para isso devemos executar um comando SQL do tipo *create table* onde faça uma tabela e utilize os mesmos atributos da tabela SAFICATE original.

3.2.1.3 Inserção na tabela

A inserção dos dados é necessária para que possamos fazer o processamento deles no processo de KDD.

Com a tabela criada, é necessário inserir os dados da tabela SAFICATE para a nova tabela sendo possível o processamento destes dados. A inserção foi feita com o comando SQL *insert into*. Com o comando executado a tabela SAFICATE2 estará

populada com todos os dados da tabela SAFICATE e prontos para o processamento dos dados.

Após a inserção dos dados é possível utilizar o comando SQL *select* para a visualização dos dados dentro Tabela 5.

Tabela 5 - Resultado Seleção

	usucodigo integer	fatcodigo integer	usacodigo smallint
1	163849	79	8
2	163860	50	3
3	163862	32	8
4	163862	37	8
5	163862	39	8

Fonte: Autoria Própria

3.2.2 Pré-processamento dos Dados

Para que seja possível utilizar os dados da tabela SAFICATE2 é necessário filtrar os dados ausentes e dados que não são uteis. Este procedimento é utilizado para facilitar a descoberta de perfis de doenças pois irá deixar a base de dados com o maior número de informação possível para ser extraída durante a fase de mineração de dados.

Nesta etapa é necessário apagar dados do prontuário eletrônico de pessoas que estão inativas, que não possuem identificação de doenças, e de pessoas que foram apenas para fazer exames gerais no SUS.

3.2.2.1 Dados ausentes

Para os dados ausentes foi feito a remoção das instâncias que não possuem doença ou estão com situação inativa e é necessário a exclusão de dados ausentes nos prontuários onde possuem doenças e os usuários estão ativos. Esta segunda remoção é feita em atributos que possuem não possuem medidas e não possuem as respostas de sim e não.

3.2.2.2 Dados Discrepantes

Com os dados ausentes excluídos, o próximo passo foi excluir dados que são discrepantes e que podem alterar o resultado do trabalho. Os dados verificados que podem possuir discrepâncias são os de altura e peso. Tal exclusão foi feita por meio de um comando SQL.

3.2.3 Dados Derivados

Os dados derivados que são necessários para este trabalho são as faixas etárias, tipos de sexo de cada indivíduo e o grupo CID. Estes dados são atributos para a geração da árvore de regras e o grupo *CID* se torna o classificador na mineração de dados.

A função para obtenção do tipo de sexo de cada usuário tem como objetivo receber uma entrada com o código de um usuário e retornar um caractere representando o sexo deste sendo as respostas possíveis M ou F, onde M corresponde ao sexo masculino e F ao feminino. O código da função pode ser visualizado na Figura 12.

Figura 12 - Código da Função para determinar o tipo de sexo

```
1  CREATE OR REPLACE FUNCTION public.tiposexo(usuario integer)
2    RETURNS char AS
3  $BODY$
4    DECLARE
5      sexo char;
6
7    BEGIN
8
9      SELECT  sextipo into sexo
10       FROM sasexos,sausuari
11       where sausuari.sexcodigo = sasexos.sexcodigo
12         and sausuari.usucodigo = usuario;
13
14     RETURN sexo;
15
16   END;
17
18
19 $BODY$
```

Fonte: (FEUSER, 2017)

Para obtenção do dado derivado grupo CID é necessário a criação de uma nova coluna na tabela SAFICATE2. A função para definir o grupo CID de cada prontuário é feita través do subgrupo do CID chamado na tabela SAFICATE2 de *cidprinci* que está presente em cada prontuário. O código da função para determinar grupo CID pode ser visualizado na Figura 13.

Figura 13 - Código da Função para determinar o grupo CID

```
1  CREATE FUNCTION public.grupocid (cidprinci character varying)
2        RETURNS character varying AS $BODY$
3  DECLARE
4        i integer;
5        grupo varchar(5);
6
7   BEGIN
8        i = cast(substr(cidprinci,2,2) as integer);
9        grupo = substr(cidprinci,1,1);
10
11       if grupo = 'A' then
12             if i < 15  then return 'A00'; end if;
13             if i between 15 and 19 then return 'A15'; end if;
14            if i between 20 and 29 then return 'A20'; end if;
15             if i between 30 and 49 then return 'A30'; end if;
16             if i between 50 and 64 then return 'A50'; end if;
17             if i between 65 and 69 then return 'A65'; end if;
18             if i between 70 and 74 then return 'A70'; end if;
19             if i between 75 and 79 then return 'A75'; end if;
20             if i between 80 and 89 then return 'A80'; end if;
21             if  89 < i  then return 'A90'; end if;
22         end if;
23
24       if grupo = 'B' then···
38       if grupo = 'C' then···
56       if grupo = 'D' then···
```

Fonte: (FEUSER, 2017)

Para a obtenção do dado derivado faixa etária será necessário a obtenção da idade de cada pessoa do prontuário eletrônico. A função para obtenção da faixa etária pode ser visualizada na Figura 12.

Figura 14 - Faixa Etária

```
CREATE FUNCTION public.faixaetaria(codigo integer)
RETURNS CHARACTER AS
$BODY$
DECLARE faixa char;
  BEGIN
select
  case
     when FLOOR(( '2016-12-31' - usudatnas )/365.25) <= 7 then '0'
     when FLOOR(( '2016-12-31' - usudatnas )/365.25) >= 8 and  FLOOR(( '2016-12-31' - usudatnas )/365.25) <= 17 then '1'
     when FLOOR(( '2016-12-31' - usudatnas )/365.25) >= 18 and FLOOR(( '2016-12-31' - usudatnas )/365.25) <= 29 then  '2'
     when FLOOR(( '2016-12-31' - usudatnas )/365.25) >= 30 and FLOOR(( '2016-12-31' - usudatnas )/365.25) <= 40 then  '3'
     when FLOOR(( '2016-12-31' - usudatnas )/365.25) >= 40 and FLOOR(( '2016-12-31' - usudatnas )/365.25) <= 60 then  '4'
     when FLOOR(( '2016-12-31' - usudatnas )/365.25) >= 61 then '5'
    END
     into faixa
   FROM sausuari
   WHERE sausuari.usucodigo = codigo;
   RETURN faixa;
   END;

$BODY$

LANGUAGE plpgsql;
```

Fonte: Autoria Própria

A função tem como objetivo principal descobrir quantos anos as pessoas possuíam no ano de 2016. Para descobrir isto é feito uma conta com a data do ano de 2016 subtraindo o valor da data de nascimento de cada pessoa. Assim é possível determinar quantos anos cada pessoa do prontuário eletrônico possuí e as classificar por faixa etária.

3.2.4 Transformação

A transformação é feita da base de dados para um arquivo CSV onde será aberto em Excel para que seja visualizado a amostragem em tabelas dinâmicas e posteriormente feita uma transformação do arquivo em um arquivo ARFF.

3.2.4.1 Seleção de Atributos

A seleção de atributos deste trabalho, conforme mencionado anteriormente na seção de seleção das colunas onde foi escolhida as colunas da tabela SAFICATE2 as quais são utilizadas para a procura de perfis em prontuários no KDD.

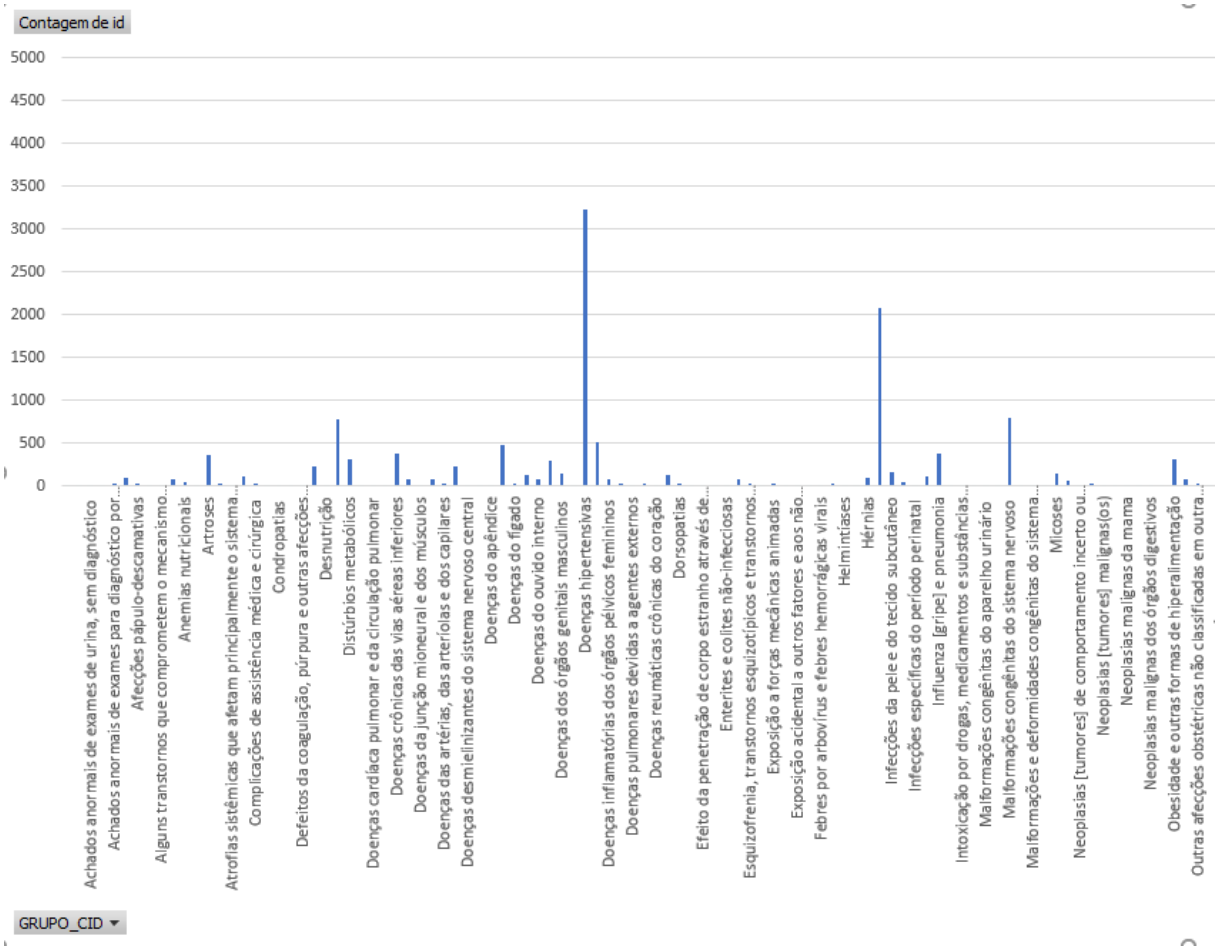
3.2.4.2 Discretização

A discretização foi realizada no próprio WEKA com seus atributos base. Foi necessária devido ao alto número de possibilidades para os atributos de altura e peso fazendo com que estes se tornassem um número total menor.

3.2.5 Amostragem

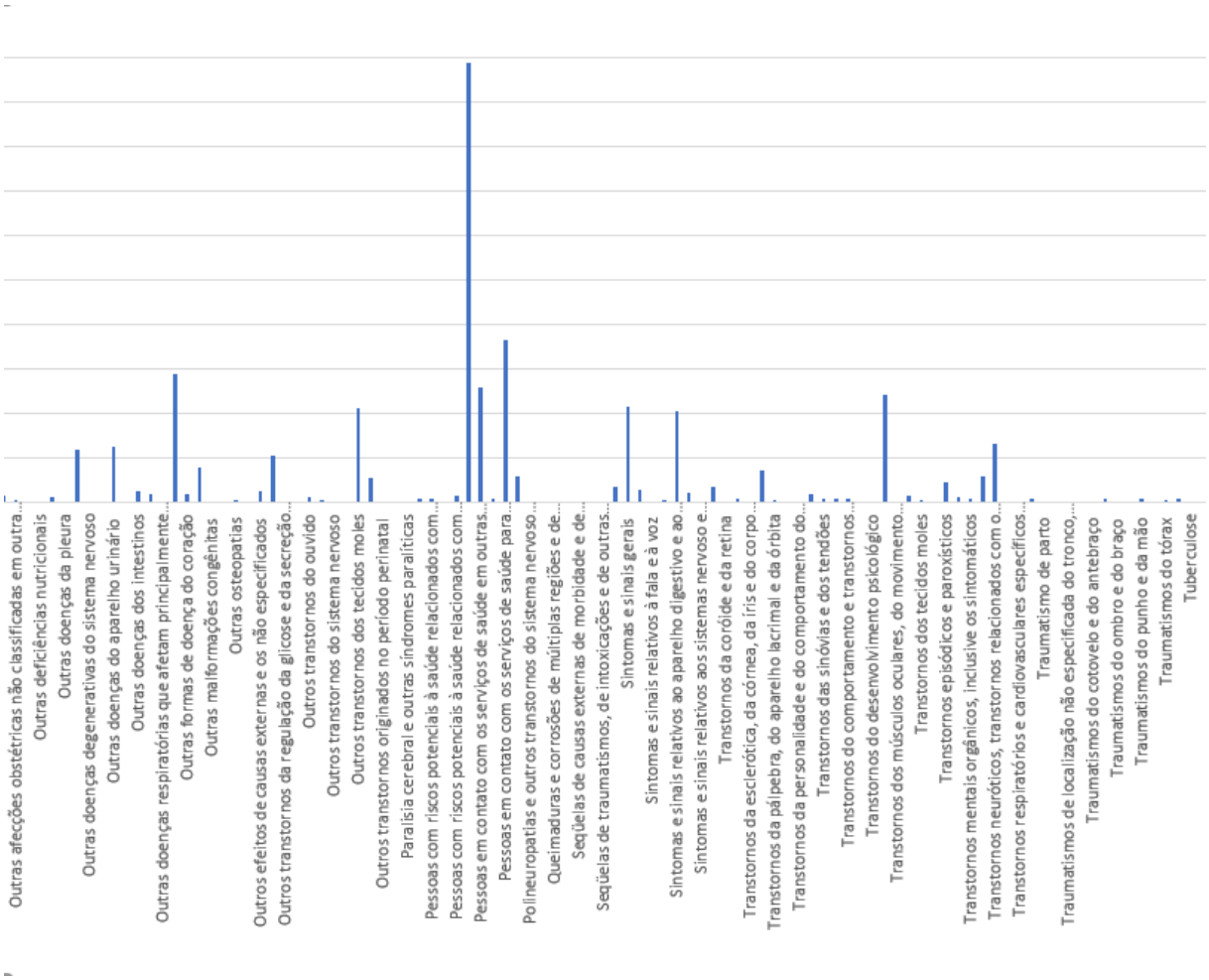
Os grupos CIDs para uma única mineração não seria possível validar a saída devido à alta discrepância entre os grupos CIDs como pode ser visto no Gráfico 1 e Gráfico 2.

Gráfico 1 - Todos *CIDS* Parte 1



Fonte: Autoria Própria

Gráfico 2 - Todos os CIDS Parte 2



Fonte: Autoria Própria

Por fim a amostragem foi dividida em dois grupos CIDs: (i) grupo de doenças relacionados a traumatismos e (ii) grupo de doenças relacionados a neoplasias.

3.2.6 Transformação

Para a criação do código em que transforma a tabela SAFICATE2 em CSV foi criado um código em linguagem PHP onde a entrada é um arquivo texto que contém uma tabela e esta tabela é transformada em comandos SQL. É possível visualizar a transformação do texto que foi lido dentro de um laço de execução onde é foi feito para a transformação da tabela em linhas de código SQL, visto na Figura 15.

Figura 15 - Função PHP

```php
<?php
$handle = @fopen("D:/TESTE/CID.txt", "r");
if ($handle) {
    while (($buffer = fgets($handle, 4096)) !== false) {

        $x = substr($buffer,0,3);
        $y = substr($buffer,4,-1);
        echo utf8_encode("WHEN CIDGRUPO='".$x."'    THEN '".$y."'<br>");
    }
    if (!feof($handle)) {
        echo "Erro: falha inexperada de fgets()\n";
    }

    fclose($handle);
}
```

Fonte: Autoria Própria

Com as linhas de códigos prontos é possível fazer a criação do código SQL em que irá selecionar os dados para que possa ser feito a mineração e obter conhecimento sobre a base de dados.

Foram realizadas três principais consultas, uma com todas as doenças do grupo CID, uma com apenas os grupos de traumatismos e uma com os grupos de neoplasias.

Para que seja feito o CSV se for para grupos específicos é necessária a criação de *views* fazendo possível escolha de colocar em CSV apenas as linhas da tabela onde possuem as doenças do grupo em específico.

3.2.6.1 *Views*

As *views* realizadas neste trabalho foram para os grupos de CIDs de neoplasias e de traumatismos. Estas podem ser visualizadas na Figura 16 e Figura 17.

Figura 16 - Código View Traumatismos

```
create view traumatismos as (
select
-- tipo sexo
CASE WHEN tiposexo(F.USUCODIGO)='M' THEN 'MASCULINO' ELSE 'FEMININO' END "Sexo",
-- faixa etaria
CASE WHEN faixaetaria(F.USUCODIGO) = '0' THEN '0 A 7' ···
-- CASE ALTURA
(SELECT CASE WHEN F.FATALTURA IS NOT NULL THEN F.fataltura END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO  ) "Altura",
-- CASE PESO
(SELECT CASE WHEN F.FATPESO > 0 THEN F.FATPESO END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO) "Peso",
--CASE "Frequenta Escola"
( SELECT CASE WHEN USU.ESUFREESC > 0  THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO  ) "Frequenta Escola",
-- CASE BENZEDEIRA
(SELECT CASE WHEN  USU.ESUFRECUR > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Frequenta Benzedeira",
-- CASE PLANO DE SAUDE
(SELECT CASE WHEN USU.ESUPLASAU > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Possui Plano de Saude",
-- CASE GESTANTE
(SELECT CASE WHEN USU.ESUGESTAN > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Gestante",
-- CASE HIPERTENSAO
(SELECT CASE WHEN USU.ESUHIPART > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO) "Hipertensao Arterial",
-- CASE DIABETES
(SELECT CASE WHEN USU.ESUDIABET > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Diabetes",
-- CASE ASMA
(SELECT CASE WHEN USU.ESUASMA   > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Asma",
-- CASE ALCOOLATRA
(SELECT CASE WHEN USU.ESUALCOOL > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Alcoolatra",
-- CASE FUMANTE
(SELECT CASE WHEN USU.ESUFUMANT > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Fumante",
-- CASE INFARTO
(SELECT CASE WHEN USU.ESUINFART > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Teve Infarto",
-- CASE AVC
(SELECT CASE WHEN  USU.ESUAVCDER > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Teve AVC/Derrame",
-- CASE BAIRROS ···
-- CASE CID  ···
-- limit 1 para que seja feito 1 pesquisa para cada pessoa
from saficate2 F
where fatdata between '01-01-2015' and '31-12-2016'
    AND f.usucodigo IN (select u.usucodigo FROM SASUSUSU u)
    group by  f.usucodigo,f.fataltura,f.fatpeso
)
```

Fonte: Autoria Própria

Figura 17 - Código View Neoplasias

```
create view neoplasias as (
select
-- tipo sexo
CASE WHEN tiposexo(F.USUCODIGO)='M' THEN 'MASCULINO' ELSE 'FEMININO' END "Sexo",
-- faixa etaria
CASE WHEN faixaetaria(F.USUCODIGO) = '0' THEN '0 A 7' ···
-- CASE ALTURA
(SELECT CASE WHEN F.FATALTURA IS NOT NULL THEN F.fataltura END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO  ) "Altura",
-- CASE PESO
(SELECT CASE WHEN F.FATPESO > 0 THEN F.FATPESO END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO) "Peso",
--CASE "Frequenta Escola"
( SELECT CASE WHEN USU.ESUFREESC > 0  THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO  ) "Frequenta Escola",
-- CASE BENZEDEIRA
(SELECT CASE WHEN  USU.ESUFRECUR > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Frequenta Benzedeira",
-- CASE PLANO DE SAUDE
(SELECT CASE WHEN USU.ESUPLASAU > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Possui Plano de Saude",
-- CASE GESTANTE
(SELECT CASE WHEN USU.ESUGESTAN > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Gestante",
-- CASE HIPERTENSAO
(SELECT CASE WHEN USU.ESUHIPART > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO) "Hipertensao Arterial",
-- CASE DIABETES
(SELECT CASE WHEN USU.ESUDIABET > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Diabetes",
-- CASE ASMA
(SELECT CASE WHEN USU.ESUASMA   > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Asma",
-- CASE ALCOOLATRA
(SELECT CASE WHEN USU.ESUALCOOL > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Alcoolatra",
-- CASE FUMANTE
(SELECT CASE WHEN USU.ESUFUMANT > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Fumante",
-- CASE INFARTO
(SELECT CASE WHEN USU.ESUINFART > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Teve Infarto",
-- CASE AVC
(SELECT CASE WHEN  USU.ESUAVCDER > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Teve AVC/Derrame",
-- CASE BAIRROS ···
-- CASE CID  ···
-- limit 1 para que seja feito 1 pesquisa para cada pessoa
from saficate2 F
where fatdata between '01-01-2015' and '31-12-2016'
    AND f.usucodigo IN (select u.usucodigo FROM SASUSUSU u)
    group by  f.usucodigo,f.fataltura,f.fatpeso
)
```

Fonte: Autoria Própria

3.2.6.2 Transformação em CSV

Para a transformação em CSV é necessário usar o comando SQL *copy* para que o resultado da consulta dentro do *copy* seja colocado em um arquivo externo. As consultas realizadas para criação dos CSVs podem ser visualizadas na Figura 18. A transformação dos dados de todos os grupos CIDs e para os grupos em específicos de neoplasias e traumatismos podem ser visualizadas na Figura 19 e Figura 20.

Figura 18 - Código Transformação para CSV todos os *cids*

```
copy (
select
-- tipo sexo
CASE WHEN tiposexo(F.USUCODIGO)='M' THEN 'MASCULINO' ELSE 'FEMININO' END "Sexo",
-- faixa etaria
CASE WHEN faixaetaria(F.USUCODIGO) = '0' THEN '0 A 7'
     WHEN faixaetaria(F.USUCODIGO) = '1' THEN '8 A 17'
     WHEN faixaetaria(F.USUCODIGO) = '2' THEN '18 A 30'
     WHEN faixaetaria(F.USUCODIGO) = '3' THEN '30 A 40'
     WHEN faixaetaria(F.USUCODIGO) = '4' THEN '40 A 60'
     WHEN faixaetaria(F.USUCODIGO) = '5' THEN 'ACIMA DE 60'
     END "Faixa etaria",
-- CASE ALTURA
(SELECT CASE WHEN F.FATALTURA IS NOT NULL THEN F.fataltura END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO  ) "Altura",
-- CASE PESO
(SELECT CASE WHEN F.FATPESO > 0 THEN F.FATPESO END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO) "Peso",
--CASE "Frequenta Escola"
( SELECT CASE WHEN USU.ESUFREESC > 0  THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO  ) "Frequenta Escola",
-- CASE BENZEDEIRA
(SELECT CASE WHEN  USU.ESUFRECUR > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Frequenta Benzedeira",
-- CASE PLANO DE SAUDE
(SELECT CASE WHEN USU.ESUPLASAU > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Possui Plano de Saude",
-- CASE GESTANTE
(SELECT CASE WHEN USU.ESUGESTAN > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Gestante",
-- CASE HIPERTENSAO
(SELECT CASE WHEN USU.ESUHIPART > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO) "Hipertensao Arterial",
-- CASE DIABETES
(SELECT CASE WHEN USU.ESUDIABET > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Diabetes",
-- CASE ASMA
(SELECT CASE WHEN USU.ESUASMA   > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Asma",
-- CASE ALCOOLATRA
(SELECT CASE WHEN USU.ESUALCOOL > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Alcoolatra",
-- CASE FUMANTE
(SELECT CASE WHEN USU.ESUFUMANT > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Fumante",
-- CASE INFARTO
(SELECT CASE WHEN USU.ESUINFART > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Teve Infarto",
-- CASE AVC
(SELECT CASE WHEN  USU.ESUAVCDER > 0 THEN 'SIM' ELSE 'NAO' END FROM SASUSUSU USU WHERE USU.usucodigo = F.USUCODIGO ) "Teve AVC/Derrame",
-- CASE BAIRROS
    (SELECT···
-- CASE TODOS CIDS
    (SELECT
      CASE
        WHEN CIDGRUPO='A00'    THEN 'Doenças Infecciosas intestinais'
        WHEN CIDGRUPO='A15'    THEN 'Tuberculose'
        WHEN CIDGRUPO='A20'    THEN 'Algumas doenças bacterianas zoonóticas'
        WHEN CIDGRUPO='A30'    THEN 'Outras doenças bacterianas'
        WHEN CIDGRUPO='A50'    THEN 'Infecções de transmissão predominantemente sexual'
        WHEN CIDGRUPO='A65'    THEN 'Outras doenças por espiroquetas'
```

Fonte: Autoria Própria

Figura 19 - Transformação para CSV Traumatismos

```
copy(
select * from traumatismos where traumatismos is not null
)
TO 'C:\Users\Public\finalTraumatismos.csv'
DELIMITER ','
CSV HEADER
```

Fonte: Autoria Própria

Figura 20 - Transformação para CSV Neoplasias

```
copy(
select * from neoplasias where neoplasias is not null
)
TO 'C:\Users\Public\finalNeoplasias.csv'
DELIMITER ','
CSV HEADER
```

Fonte: Autoria Própria

3.2.6.3CSV para ARFF

A transformação de CSV para ARFF é feita na ferramenta WEKA. Basta selecionar em *tools* a opção ARFF *viewer*, depois disso é feita a escolha dos arquivos CSV que serão transformados em arquivos ARFF.

3.2.7 Mineração de Dados

Para a mineração de dados foi utilizada a ferramenta WEKA, onde foram executados os algoritmos C4.5, *Bagging* e *Boosting*. Os algoritmos foram feitos utilizando do padrão C4.5. Para a execução no WEKA também é escolhido o método de testes *Cross-Validation* para a execução dos três algoritmos.

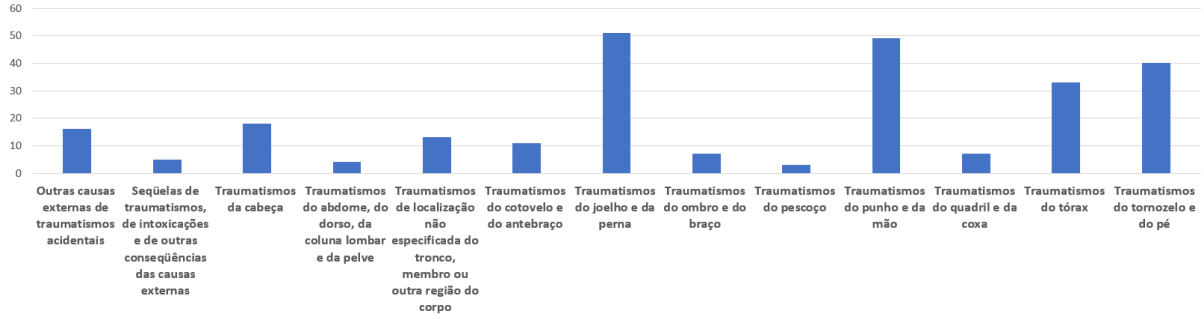
4 RESULTADOS

Neste capítulo será tratada a aplicação da mineração de dados nos dois grupos de doenças escolhidos durante o trabalho. O capítulo é dividido em: (i) Mineração do Grupo de Doenças de Neoplasias, (ii) Mineração do Grupo de Doenças de Traumatismos.

4.1 MINERAÇÃO DO GRUPO DE DOENÇAS DE NEOPLASIAS

O grupo de neoplasias consistem nas doenças do grupo CID de C00 até D37, que somam um total de 10 tipos de doenças. Estas doenças têm seus respectivos volumes de instancias que somados dão 119 e o grupo pode ser visualizados no Gráfico 3.

Gráfico 3 - Instâncias de Neoplasias



Fonte: Autoria Própria

Ao executar o C4.5 é criada uma arvore de decisão de 55 ramos e 64 de tamanho. É possível visualizar um trecho desta árvore na Figura 21.

Figura 21 - Trecho das Regras da Árvore de Decisão dos Neoplasias

```
BAIRRO = PINHEIRINHO: Neoplasias [tumores] in situ (11.0/1.0)
BAIRRO = SANTA TEREZINHA: Neoplasias [tumores] in situ (4.0)
BAIRRO = ALVORADA
|   Frequenta Escola = NAO: Neoplasias [tumores] benignas(os) (17.0/3.0)
|   Frequenta Escola = SIM: Neoplasias malignas do aparelho respiratï¿½rio e dos ï¿½rgï¿½os intratorï¿½cicos (7.0)
BAIRRO = SAO FRANCISCO: Neoplasias [tumores] benignas(os) (11.0)
BAIRRO = PLANALTO
|   Faixa etaria = 40 A 60: Neoplasias [tumores] benignas(os) (4.0/1.0)
|   Faixa etaria = 0 A 7: Neoplasias [tumores] in situ (0.0)
|   Faixa etaria = 30 A 40: Neoplasias [tumores] in situ (0.0)
|   Faixa etaria = ACIMA DE 60: Neoplasias [tumores] in situ (4.0/1.0)
|   Faixa etaria = 18 A 30: Neoplasias [tumores] in situ (0.0)
BAIRRO = MORUMBI
|   Faixa etaria = 40 A 60: Neoplasias malignas da mama (2.0)
|   Faixa etaria = 0 A 7: Neoplasias [tumores] benignas(os) (0.0)
|   Faixa etaria = 30 A 40: Neoplasias [tumores] benignas(os) (2.0)
|   Faixa etaria = ACIMA DE 60: Neoplasias [tumores] benignas(os) (0.0)
|   Faixa etaria = 18 A 30: Neoplasias [tumores] benignas(os) (0.0)
BAIRRO = VENEZA: Neoplasias [tumores] benignas(os) (2.0)
BAIRRO = VILA VERDE: Neoplasias [tumores] benignas(os) (2.0/1.0)
```

Fonte: Autoria Própria

A Figura 21 ilustra um trecho da árvore de decisão gerada pelo algoritmo J48, onde o primeiro ramo é bairro, segundo ramo é a frequência escolar, seguido pela faixa etária e o ultimo ramo possível sendo a altura. Mostrando assim a ordem de maior ganho de informação dentro da árvore gerada para as doenças relacionadas a neoplasias.

Com a árvore de decisão obtida é possível que seja feito uma análise por pessoas da área da saúde para um eventual sistema de prevenção para o principal tipo de doença classificado em cada bairro ou doenças.

Para a análise da árvore pode ser feita a porcentagem de classificação para cada bairro. Esta porcentagem é feita com o número de classificação de cada ramo da árvore dividido pelo número total de instancias.

- Bairro Pinheirinho: 11 / 119 = 9,24%;
- Bairro Santa Terezinha: 4 / 119 = 3,3%;
- Bairro Alvorada Não Frequenta Escola: 17 / 119 = 28.5%;
- Bairro Alvorada Frequenta escola: 7 / 119 = 5,8%;
- Bairro São Francisco: 11 / 119 = 9.24%.

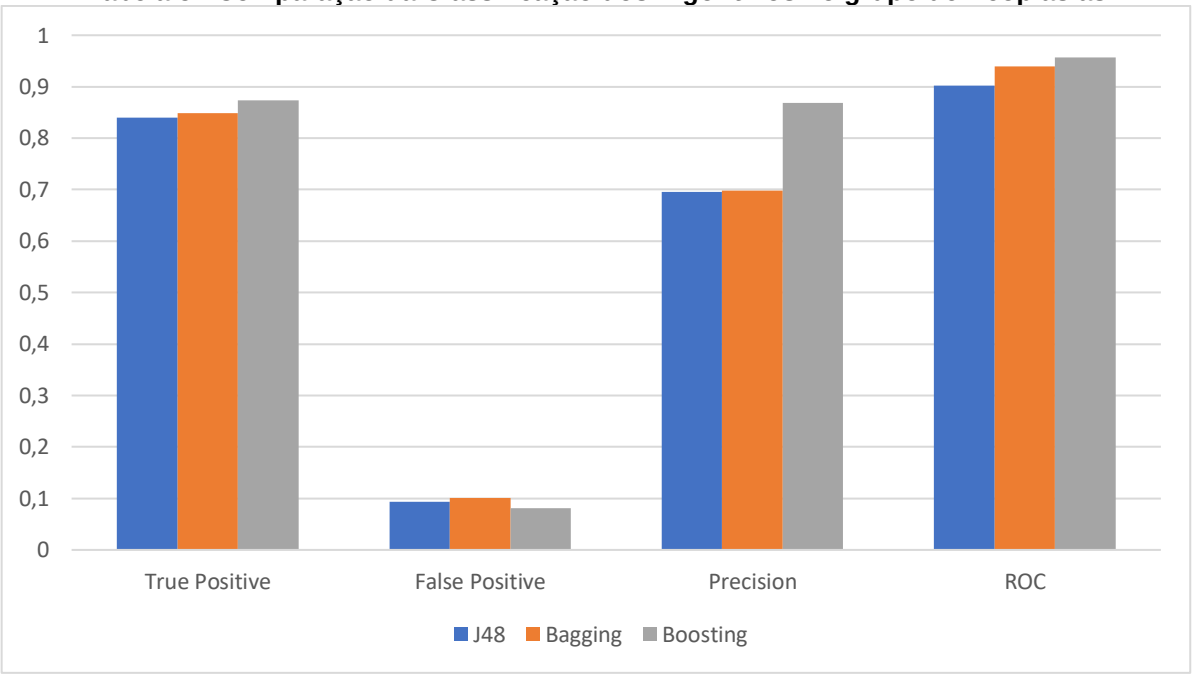
Com essas probabilidades analisadas é possível determinar os bairros com maiores taxas de classificação como a regra no bairro Alvorada com um total de 34,3% dos dados classificados. No bairro do Alvorada também foi possível constatar o perfil das pessoas que sofrem com Neoplasias nesta região, destas aquelas as quais

frequentam a escola apresentam uma classificação de neoplasia respiratória e as que não frequentam a escola em geral apresentam neoplasias benignas.

Com essas regras é possível definir os primeiros bairros para se fazer uma campainha de conscientização e medidas preventivas.

Com a aplicação de outros dois algoritmos para este grupo de doenças é possível perceber uma melhoria na classificação dos atributos. Esta melhoria pode ser visualizada na Tabela 5.

Tabela 5 - Comparação da Classificação dos Algoritmos no grupo de Neoplasias



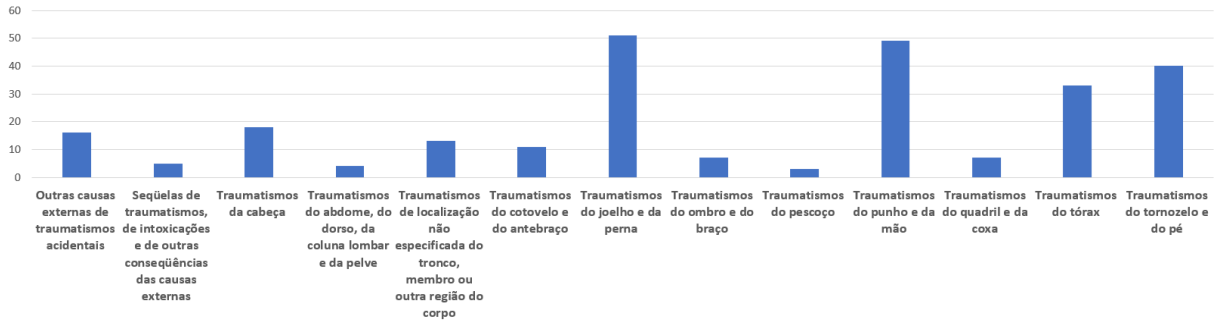
Fonte: Autoria Própria

A Tabela 5 mostra claramente a melhora da classificação quando aplicamos o *Bagging* e o *Boosting* para os grupos de doenças de neoplasias. Baseado na figura é possível determinar que o algoritmo *Boosting* com C4.5 possui a melhor classificação para este grupo de doenças, pois ele possui a maior precisão dentre os três e a menor taxa de instâncias classificadas erroneamente.

4.2 MINERAÇÃO DO GRUPO DE DOENÇAS DE TRAUMATISMOS

O grupo de doenças de traumatismos consiste nos grupos CIDs de S00 a T08, totalizando onze grupos, estes grupos podem ser visualizados no Gráfico 4 com seus respectivos volumes que somados totalizam 236.

Gráfico 4 - Instâncias de Traumatismos



Fonte: Autoria própria.

O Gráfico 4 mostra a comparação de volume entre os grupos de doenças relacionadas a traumatismos, concluindo que os maiores números de traumatismos na região foram de joelhos, pernas, punhos e mãos.

Ao executar o algoritmo C4.5 é gerado uma árvore de decisão de tamanho total de 203 ramos com 173 nós possíveis. Um trecho desta árvore pode ser visto na Figura 22.

Figura 22 - Trecho das Regras da Árvore de Decisão dos Traumatismos

```
BAIRRO = MORUMBI
|   Faixa etaria = 0 A 7: Traumatismos do tornozelo e do pï¿½  (1.0)
|   Faixa etaria = 40 A 60
|   |   Sexo = MASCULINO: Traumatismos do joelho e da perna  (3.0)
|   |   Sexo = FEMININO: Traumatismos do ombro e do braï¿½o  (2.0)
|   Faixa etaria = ACIMA DE 60: Traumatismos do joelho e da perna  (0.0)
|   Faixa etaria = 30 A 40: Traumatismos do joelho e da perna  (0.0)
|   Faixa etaria = 18 A 30: Traumatismos do punho e da mï¿½o  (3.0)
|   Faixa etaria = 8 A 17: Traumatismos do joelho e da perna  (0.0)
BAIRRO = PINHEIRINHO
|   Faixa etaria = 0 A 7: Traumatismos do joelho e da perna  (0.0)
|   Faixa etaria = 40 A 60
|   |   Hipertensao Arterial = NAO: Traumatismos do joelho e da perna  (3
|   |   Hipertensao Arterial = SIM: Traumatismos do cotovelo e do antebra
|   Faixa etaria = ACIMA DE 60: Traumatismos do joelho e da perna  (3.0)
|   Faixa etaria = 30 A 40: Traumatismos do joelho e da perna  (0.0)
|   Faixa etaria = 18 A 30: Traumatismos do tornozelo e do pï¿½  (6.0/3.0
|   Faixa etaria = 8 A 17: Traumatismos do joelho e da perna  (0.0)
BAIRRO = SAO ROQUE
|   Sexo = MASCULINO: Traumatismos do tornozelo e do pï¿½  (3.0/2.0)
|   Sexo = FEMININO
|   |   Hipertensao Arterial = NAO: Traumatismos do tï¿½rax  (3.0)
|   |   Hipertensao Arterial = SIM: Traumatismos do joelho e da perna  (2
BAIRRO = PLANALTO
|   Faixa etaria = 0 A 7: Traumatismos do tornozelo e do pï¿½  (2.0/1.0)
|   Faixa etaria = 40 A 60
|   |   Sexo = MASCULINO: Traumatismos do tï¿½rax  (10.0/7.0)
|   |   Sexo = FEMININO: Traumatismos do joelho e da perna  (3.0/1.0)
|   Faixa etaria = ACIMA DE 60: Traumatismos do tï¿½rax  (3.0/1.0)
|   Faixa etaria = 30 A 40
|   |   Sexo = MASCULINO: Traumatismos do tornozelo e do pï¿½  (5.0/3.0)
|   |   Sexo = FEMININO: Traumatismos do punho e da mï¿½o  (4.0)
|   Faixa etaria = 18 A 30: Traumatismos do punho e da mï¿½o  (11.0/4.0)
|   Faixa etaria = 8 A 17: Traumatismos do tornozelo e do pï¿½  (5.0/1.0)
```

Fonte: Autoria Própria

Para os traumatismos é possível perceber que a árvore tem seu primeiro ramo de regras no atributo bairro seguido por faixa etária e sexo. Com isto pode-se perceber que alguns dos atributos escolhidos durante a etapa de escolha de atributos não possuem um ganho de informação relevante para estar presente na árvore de decisão.

Assim como em neoplasias esta análise pode ajudar pessoas que trabalham na área da saúde a melhorar o tratamento em algumas regiões.

Para a análise da árvore da árvore do grupo de doenças de traumatismos serão comparados alguns dos mesmos bairros em que foi realizada a análise no grupo de neoplasias. A seguir é indicada a porcentagem de classificação para cada bairro:

- Bairro Pinheirinho: 14/236 = 5,93 %
- Bairro Morumbi: 9/236 = 3,8%
- Bairro Santa Terezinha: 11/236 = 4,6%
- Bairro Alvorada: 26/236 = 11%
- Bairro Planalto: 43 / 236 = 18,2%

Nesta análise também é possível detalhar mais sobre o perfil das pessoas que moram nestes bairros mostrados as porcentagens acima.

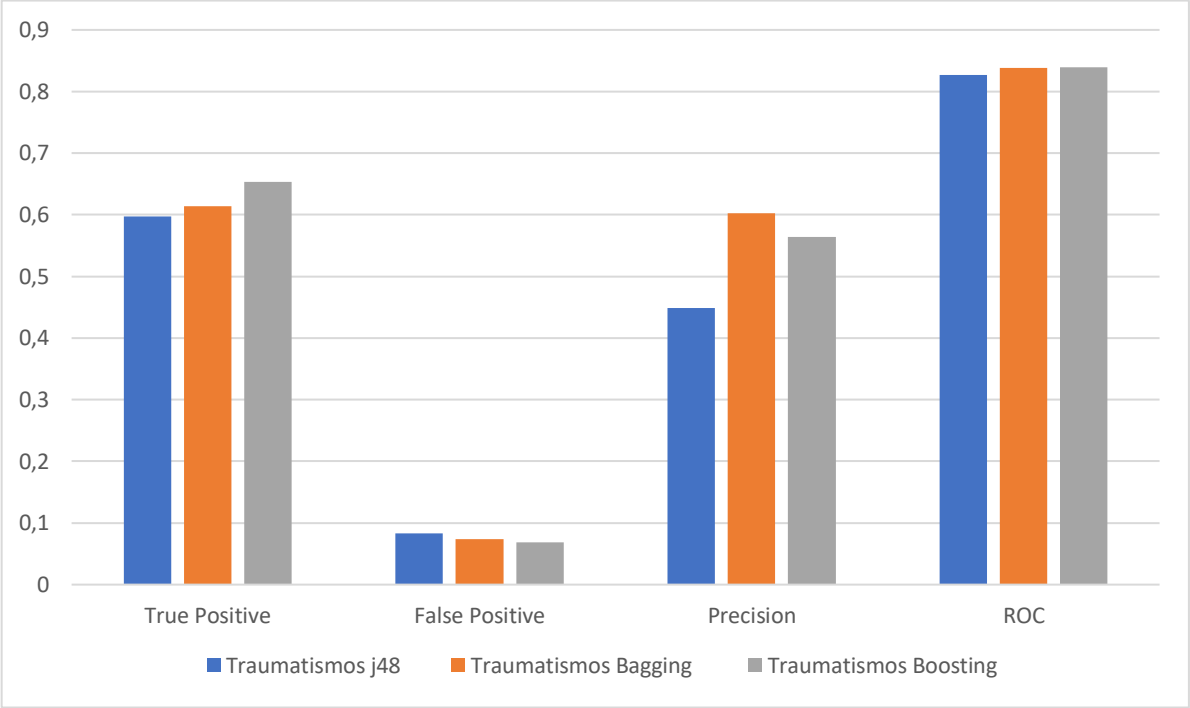
No bairro Pinheirinho os ramos se dividem em faixa etária onde os maiores índices de classificação ocorreu nas faixas etárias de 40 a 60 anos de idade com traumatismos de cotovelo e antebraço para quem possui hipertensão arterial e traumatismos de joelho e perna para quem não possui. Para as pessoas de faixa etária de 18 a 30 anos de idade com traumatismos de tornozelo e pé.

Para o bairro Morumbi é possível visualizar que: (i) pessoas de faixa etária de 0 a 7 anos tem uma classificação de traumatismos de tornozelo e pé, (ii) as pessoas de faixa etária de 40 a 60 anos se forem homens possuem traumatismos de joelho e perna e se mulheres traumatismos de ombro e braço, (iii) pessoas de faixa etária entre 18 a 30 anos possuem a classificação nos traumatismos de punho e perna.

O bairro Planalto possui o maior número de classificações. Verifica-se que o maior número de casos de classificação ocorreu com pessoas de faixa etária entre 40 a 60 anos, 8 destes casos ocorreram com pessoas do sexo masculino para traumatismos de tórax, já para as mulheres a classificação foi em traumatismos de joelho e perna. O segundo maior número de classificações neste bairro foi para faixa etária de 18 a 30 anos de idade para traumatismos de punhos e mãos.

Com a execução dos demais algoritmos de mineração foi possível gerar a Tabela 6.

Tabela 6 - Comparação J48, Bagging e Boosting para Traumatismos



Fonte: Autoria Própria

Por meio da Tabela 6 é possível observar que ao serem executados os algoritmos de modo geral são reduzidos os números de instâncias falsamente classificadas. O *Bagging* possui uma curva crescente quando comparado ao C4.5 e apresenta melhor precisão que o *Boosting.* O *Boosting* por sua vez apresenta mais dados verdadeiros classificados do que os outros dois classificadores.

5 CONCLUSÃO

A mineração de dados é cada vez mais necessária com a alta crescente dos bancos de dados no mundo, pois cada vez mais temos bancos de dados os quais podem conter conhecimento não explorado que pode nos auxiliar na melhoria da qualidade de vida, este é o caso dos bancos de dados voltados a área da saúde.

Algoritmos de aprendizagem de máquina, como por exemplo o C4.5 são utilizados com sucesso em bancos de dados distintos e quando estes aplicados a variantes diferentes como *Bagging* e *Boosting* podem melhorar significativamente o processo de classificação como pode ser visto neste trabalho.

O objetivo deste trabalho foi a analisar padrões de perfil da base de dados do sistema de saúde. Dentro destes dados foram aplicados:

(i) Seleção dos Dados;
(ii) Pré-processamento dos Dados;
(iii) Amostragem;
(iv) Transformação;
(v) Mineração de Dados com algoritmos C4.5, *Bagging* e *Boosting*; e
(vi) Resultados.

Ao finalizar a aplicação das etapas do KDD para análise de perfis de doenças e dividir a amostra em dois grupos de doenças, um relacionado a neoplasias e o outro relacionado a traumatismos, foi possível determinar na análise que para o grupo de neoplasias os principais atributos foram bairros, frequência escolar, faixa etária e altura. Por outro lado, quando se trata do grupo de doenças relacionadas a traumatismos os principais atributos encontrados foram os bairros, faixa etária, altura, peso, sexo, faixa etária, diabetes, hipertensão e fumante. Podendo assim ter um perfil base para uma análise de um profissional da área para concluir se a descoberta de conhecimento deste trabalho.

5.1 LIMITAÇÕES DO TRABALHO

As principais limitações deste trabalho podem ser resumidas em:

(i) Falta de um profissional da saúde para analisar os dados resultantes;
(ii) Grande número de prontuários com dados ausentes.

5.2 TRABALHOS FUTUROS

Para trabalhos futuros é possível a validação dos resultados obtidos por profissionais da saúde para definir a veracidade das regras encontradas, também a escolha de novos atributos para criação de regras diferentes. Assim podendo finalizar o ciclo do KDD para a descoberta do conhecimento. Este trabalho também pode ser aplicado em prontuários eletrônicos de hospitais, ambulatórios entre outros.

REFERÊNCIAS

AGRAWAL, R.; SKIRANT R. **Fast Algorithms for Mining Associations Rules**. ACM Digital Library, San Francisco (CA), p.487- 499, set. 1994.

ALVES, L. **Prontuário Eletrônico x Prontuário no papel.** Acessado em: http://meuprontuario.net/prontuario-eletronico-x-prontuario-papel-qual-e-o-melhor/ disponível em: 24/06/2018.

ANICETO, M. **Classificadores Ensemble, tipos Bagging e Boosting.** Postado 27 setembro,2017. Disponível em: https://lamfo-unb.github.io/2017/09/27/BaggingVsBoosting/.

BORGES, A.P. **Descoberta de Regras de Condução de Locomotivas.** PUCPR-Pontifícia Universidade Católica do Paraná. Maio, 2008.

BREIMAN, L. 1996. **Bagging predictors. Machine learning,** 24(2). 1996, pp. 123-140.

CARVALHO, D.R; ESCOBAR L.F.A.; TSUNODA D. **Pontos de Atenão para o uso da mineração de dados na saúde.**

CARVALHO, R. **Prontuário e registro de enfermagem.** Acessado em: http://www.ebah.com.br/content/ABAAAASqAAG/prontuario-registro-enfermagem#. Disponível em:24/06/2018.

DICIO. **Dicionário Online de Português.** Acessado em: https://www.dicio.com.br disponível em 24/06/2018.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, American Association for Artificial Intelligence, Boston, 1996.

FELIX, L. **Data Mining No Processo de Extração de Conhecimento de Bases de dados.** USP, São Carlos (SP),1998.

FEUSER, R. **Mineração de Dados com Regras de Associação Aplicada em Dados de Unidade de Saúde de Pronto Atendimento**. UTFPR, Pato Branco (PR), 2017.

GALVÃO, N. D. **Aplicação da Mineração de Dados em Bancos da Segurança e Saúde Pública em Acidentes de Transporte.** UNIFESP, São Paulo, 2009.

GARCIA-LAENCINA, P.J.; ABREU, P.H.; ABREU, M.H.; AFONOSO, N. **Missing Data Imputation on the 5-year Survival Prediction of Breast Cancer Patients with Unknown Discrete Values.** Journal Computers in Biology and Medicine, vol.59, apr. 2015.

Gray, Jim & Andreas R. **Distributed Transaction Processing: Concepts and Techniques.** Morgan Kaufmann, 1993.

HALAMKA, J.D. **Heath Information Technology: Shaw We Wait for the Evidence? Annals of Internal Medicine,** maio, 2006.

KRYSZTOF, J.C. **Uniqueness of Medical Data Mining**. Artificial Intelligence in Medicine, mar, 2002.

LANTZ, B. **MACHINE LEARNING WITH R. PACKT.** PACKT PUBLISHING, 2013.

MARCELO, L. S, **Prontuário Eletrônico do Paciente.**, 10º AUDHOSP – Congresso Nacional de Auditoria em Saúde e Qualidade e da Assistência Hospitalar, 2011.

MITCHELL, T, **Machine Learning.** New York: mcGraw-Hill, 1997.

MONTENEGRO, M.R. **Bootstrap.** Postado 28 junho,2017. Disponível em: https://lamfo-unb.github.io/2017/06/28/Bootstrap/.

OLIVEIRA, C. C**. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Goiás, 2009

PostgreSQL, Disponível em https://www.postgreSQL.org.

QUINLAN, J. R. **C4.5 Programs for Machine Learning,** Elsevier. 1993.

QUINLAN, J. R. 1996. **Improved Use of Continuous Attributes in C4.5**. Journal of Artificial Intelligence Research. 1996, Vol. IV.

QUONIAN, L, Tarapanoff K. **Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil.** Ci Inf 2001.

RISHARD, S. DICK ELAINE B. STEEN. **The Computer-Based Patient Record An Essential Technology for Health Care.** Institute Of Medice 1991. Disponível em: https://www.nap.edu/read/18459/chapter/1

SILLA, C. N., KAESTNER, C. e KOERICH, A. 2005. **Classificação Automática de Gêneros Musicais Utilizando Métodos de Bagging e Boosting.** SBCM - Simpósio Brasileiro de Computação Musical. Outubro de 2005. Disponível em: https://www.researchgate.net/publication/239928986_Classificacao_Automatica_de_Generos_Musicais_Utilizando_Metodos_de_Bagging_e_Boosting.

SILVA, Fábia Gama; TAVARES-NETO, José**. Avaliação dos prontuários médicos de hospitais de ensino do Brasil**. Rev. bras. educ. med., Rio de Janeiro. 31, n. 2, p. 113-126.Aug.2007 Disponível em:<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-55022007000200002&lng=en&nrm=iso>. Acesso em: 24/06/2018.

THOMAZ, M. **TUDO QUE VOCÊ PRECISA SABER SOBRE PRONTUÁRIOS ELETRÔNICOS.** Disponível em: https://blog.iclinic.com.br/tudo-sobre-prontuario-eletronico/ acessado em: 24/06/2018.

TRINDADE, C M et al. **Technology in health: knowledge discovery in public health databases: study of viral hepatitis in the state of Paraná, Brazil.** Iberoamerican Journal of Applied Computing, Ponta Grossa, v. 2, n. 2, 2012.

Vianna, S. **MINERAÇÃO DE DADOS: Uma revisão da literatura em Administração.** Volume 8, Número 2, Juiz de Fora, dezembro 2017.

VILARINHO R.A. **Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros**. UFOP – Universidade Federal de Ouro Preto. Março, 2017.

WEKA, Waikato Environment for Knowledge Analysis. Disponível em *https://www.cs.waikato.ac.nz/ml/weka/index.html.*

WITTEN, I.H. **Data Mining Practical Machine Learning Tools and Techniques**. 3 ed., 2011.