UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA BIOMÉDICA (PPGEB)

WILLIAM HAMILTON DOS SANTOS

ESTUDO DA BASE DE DADOS ABERTOS E-SAÚDE DA PREFEITURA DE CURITIBA USANDO TÉCNICAS DE MINERAÇÃO DE DADOS

DISSERTAÇÃO

CURITIBA

2018

WILLIAM HAMILTON DOS SANTOS

ESTUDO DA BASE DE DADOS ABERTOS E-SAÚDE DA PREFEITURA DE CURITIBA USANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Engenharia Biomédica, do Programa de Pós-Graduação em Engenharia Biomédica da Universidade Tecnológica Federal do Paraná.

Área de Concentração: Engenharia Biomédica

Orientador: Prof. Dr. Gilson Yukio Sato

CURITIBA

2018

Dados Internacionais de Catalogação na Publicação

Santos, William Hamilton dos
Estudo da base de dados abertos E-Saúde da prefeitura de Curitiba usando técnicas de mineração de dados [recurso eletrônico] / William Hamilton dos Santos.-- 2019.
1 arquivo texto (83 f.) : PDF ; 1,31 MB.

Modo de acesso: World Wide Web.
Título extraído da tela de título (visualizado em 23 out. 2019).
Texto em português com resumo em inglês
Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Engenharia Biomédica, Curitiba, 2018
Bibliografia: f. 71-78

1. Engenharia biomédica - Dissertações. 2. Mineração de dados (Computação). 3. Banco de dados - Curitiba (PR). 4. Sistemas de recuperação da informação - Medicina. 5. Medicina - Curitiba (PR) - Processamento de dados. 6. MySQL (Recurso eletrônico). 7. Weka (Programa de computador). I. Sato, Gilson Yukio. II. Universidade Tecnológica Federal do Paraná - Programa de Pós-graduação em Engenharia Biomédica. III. Título.

CDD: Ed. 23 – 610.28

Biblioteca Central da UTFPR, Câmpus Curitiba
Bibliotecário: Adriano Lopes CRB-9/1429

Universidade Tecnológica Federal do Paraná

TERMO DE APROVAÇÃO DE DISSERTAÇÃO Nº 115

A Dissertação de Mestrado intitulada “ESTUDO DA BASE DE DADOS ABERTOS E-SAÚDE DA PREFEITURA DE CURITIBA USANDO TÉCNICAS DE MINERAÇÃO DE DADOS”, defendida em sessão pública pelo candidato **William Hamilton dos Santos**, no dia 10 de dezembro de 2018, foi julgada para a obtenção do título de Mestre em Ciências, área de concentração Engenharia Biomédica, e aprovada em sua forma final, pelo Programa de Pós-Graduação em Engenharia Biomédica.

Banca examinadora:
Prof. Dr. Gilson Yukio Sato, UTFPR – UTFPR

Profa. Dra. Raquel Kolitski Stasiu – UTFPR

Profa. Dra. Deborah Ribeiro Carvalho – PUCPR

Curitiba, 10 de dezembro de 2018.
Carimbo e Assinatura do(a) Coordenador(a) do Programa

A via original deste documento encontra-se arquivada na Secretaria do Programa, contendo a assinatura da Coordenação após a entrega da versão corrigida do trabalho.

Dedico este trabalho aos meus pais Hamilton e Maria e
a minha esposa Egislaine e
aos meus filhos Guilherme e Gabriel

AGRADECIMENTOS

Agradeço, primeiramente a Deus por tudo o que tenho e que sou. Aos meus pais, Hamilton e Maria Aparecida, pela criação, educação, apoio e incentivo.

Um agradecimento mais que especial a minha esposa, Egislaine, pelo apoio, pelo incentivo, pela paciência e compreensão pelo tempo dedicado aos estudos. Seu apoio foi decisivo para finalizar mais esta etapa de nossas vidas.

Aos meus filhos Guilherme e Gabriel, que sendo crianças pequenas e necessitando de atenção, me recebiam com carinho mesmo tendo que ceder muitos momentos de brincadeiras para os estudos do papai.

Ao meu orientador, professor Gilson Yukio Sato, pelo constante apoio, paciência e disposição na condução de todo o mestrado, especialmente pelas longas sessões de orientação.

A professora Leandra, pelas valiosas contribuições desde os primeiros instantes no curso. Ao professor Bertoldo pelas aulas inspiradoras nas várias disciplinas ministradas.

Aos professores Joaquim Miguel Maia, Luciano Scandelari e Vicente Machado Neto, pelos relevantes conhecimentos transmitidos em suas disciplinas que completaram o rol de disciplinas cursadas.

Ao casal Josi e Ton, servidores da secretaria de saúde de Curitiba que tantas vezes explicaram a dinâmica de funcionamento das unidades do sistema de saúde do município.

As gerentes das Unidades de Saúdes Érico Veríssimo e Tapajós, Carla e Alzira, que nas várias conversas forneceram o caminho para aplicar na prática toda a estrutura teórica desenvolvida.

A prefeitura de Curitiba por disponibilizar a base de dados E-Saúde, seguindo sua política de dados abertos, principal motivo de estudos deste mestrado.

As chefias da DIRGTI, setor onde trabalho no apoio dado, aos colegas de trabalho que deram opiniões sobre os diversos aspectos do trabalho desenvolvido. E a todos que direta ou indiretamente contribuíram para a realização deste sonho.

RESUMO

SANTOS, William Hamilton dos. Estudo da Base de Dados Abertos E-Saúde da Prefeitura de Curitiba usando Técnicas de Mineração de Dados. 2018. 115ª Dissertação – Programa de Pós Graduação em Engenharia Biomédica, Universidade Tecnológica Federal do Paraná. Curitiba, 2018.

Neste trabalho, foi estudada a possibilidade de explorar a base de dados pública do sistema de saúde da cidade de Curitiba, o E-Saúde usando mineração de dados, No E-Saúde são encontrados dados relativos aos atendimentos médicos realizados nos pacientes. Foram aplicadas técnicas de mineração de dados, dentro de um Processo de Descoberta de Conhecimento, nas instâncias da base de dados E-Saúde, para estudar a possibilidade de identificar padrões e correlações na forma de regras textuais capazes de servir de suporte aos gestores das Unidades de Saúde do sistema de Saúde da cidade de Curitiba. O Processo de Descoberta de Conhecimento em Bases de Dados, ou Knowledge Discovery in Databases (KDD), foi aplicado ao E-Saúde. Contudo, em função das características dessa base de dados foi necessário adaptar as fases do KDD. Os arquivos constituintes do E-Saúde, disponibilizados no Portal de Dados abertos de Curitiba, foram importados para um banco de Dados MySQL, no qual passaram por um processo de limpeza, transformação, seleção e validação dos dados. Para a fase da mineração dos aplicada a tarefa de classificação, com uso de dois algoritmos: um indutor de Árvore de Decisão e outro baseado em Regras. Os algoritmos foram aplicados em todas as instâncias referentes aos primeiro trimestre do ano de 2017 em dois Distritos Sanitários da cidade. Nos experimentos foi utilizado o Weka, software de uso livre, que contém uma coleção de algoritmos já prontos para uso. Foram aplicados os classificadores J48 e JRip no atributo (da base de dados) *Solicitação de Exames*. Este atributo é um dos que definem melhor a resolutividade de uma consulta. A resolutividade está relacionada com o fato do paciente, ter tido (ou não) seu caso resolvido. Os modelos obtidos podem ser considerados válidos, pois as acurácias ficaram em torno de 74%, indicando a aplicabilidade do processo de mineração dos dados na E-Saúde. Os resultados indicam que a E-Saúde pode ser explorada para obtenção de conhecimento potencialmente útil para a tomada de decisão de um gestor de uma Unidade Básica de Saúde.

Palavras-chave: Mineração de Dados, Base de Dados Pública, Árvore de Decisão J48 e Regras JRip.

ABSTRACT

SANTOS, William Hamilton dos. Study of the Curitiba Open Database E-Saúde using Data Mining Techniques. **Number 115**. Dissertação – Programa de Pós Graduação em Engenharia Biomédica, Federal University of Technology - Paraná. Curitiba, 2018.

In this study, I analysed the possibility to apply data mining to the public database containing data provided by the health TI system of the city of Curitiba, the E-Saúde.The E-Saúde contains data related to the medical care provided by the city. I applied data mining techniques, within a Knowledge Discovery Process, in the instances of the E-Saúde database, to study the possibility to find patterns and correlations, in the form of textual rules that could be used by managers of the Health Units to take decisions. In order to reach this objective, the Knowledge Discovery in Databases (KDD) process was applied to the database. However, because the characteristics of the database, it was necessary to adapt the KDD. The E-Saúde files made available by Curitiba's Open Data Portal were imported into a MySQL database, in which they were cleaned, processed, selected and validated. For the data mining phase, the classification task was applied using two algorithms, a Decision Tree inductor and another based on Rules. The algorithms were applied on all instances of the first quarter of 2017 in two city Districts. In the experiments I used the Weka, a free software, which contains a collection of algorithms ready for use. The Decision Tree algorithm used was J48, and the algorithm based on Rules applied was the JRip. The J48 and JRip classifiers were applied in the attribute *Solicitação de Exames*. This attribute is one of those that best define the resolutivity of a medical appointment. The resolutivity is related to if the patient had (or had not) his case solved. The obtained models can be considered valid, since the accuracy obtained were around 74%, indicating the applicability of the data mining process in the E-Saúde database. The results indicate that the E-Saúde can be explored to obtain knowledge potentially useful for the decision make process preformed by the managers of Basic Health Units.

Keywords: Data Mining. Public Databases. Decision Tree J48 and Rules JRip.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS E ACRÔNIMOS

ASCII	*American Standard Code for Information Interchange*
ARFF	*Attribute-Relation File Format*
CBO	Classificação Brasileira de Ocupações
CID	Classificação Internacional de Doenças
CSV	*Comma Separated Values*
ESF	Estratégia em Saúde da Família
KDD	*Knowledge Discovery in Databases*
RIPPER	*Repeated Incremental Pruning to Produce Error Reduction*
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	*Structured Query Language*
SSD	*Solid-State Drive*
SUS	Sistema Único de Saúde
UBS	Unidade Básica de Saúde
UPA	Unidade de Pronto Atendimento
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1. INTRODUÇÃO

No Brasil, governos de todas as esferas, federal, estadual e municipal disponibilizam portais com bases de dados públicas que podem ser descarregadas e estudadas livremente pelos mais diversos agentes da sociedade.

Uma das bases de dados abertos referentes à saúde mais conhecidas provém do Departamento de Informática do Sistema Único de Saúde do Governo Federal, o DATASUS. Tal departamento é um órgão da Secretaria de Gestão Estratégica e Participativa do Ministério da Saúde com a responsabilidade de coletar, processar e disseminar informações sobre saúde (DATASUS, 2018).

Da mesma forma que o Governo Federal, Estados e Municípios possuem políticas de dados abertos relacionados à saúde. A Prefeitura Municipal de Curitiba (PMC), por exemplo, disponibiliza bases de dados abertos que abrangem áreas como: Administração Pública, Recursos Humanos, Cultura, Habitação, Segurança, Abastecimento, Transporte, Pesquisa e Planejamento e Saúde (PORTAL, 2018).

Os dados relacionados à saúde da PMC são extraídos do sistema de informática E-Saúde. Este sistema viabiliza o registro dos atendimentos prestados pela Secretaria Municipal de Saúde (SMS) de Curitiba em sua rede de atenção (DADOS, 2018). A rede é composta por Unidades Básicas de Saúde (UBS), Unidades de Pronto Atendimento (UPA) e Centros de Especialidades Médicas e Odontológicas. Os dados disponibilizados para consulta referem-se ao perfil de atendimento dos profissionais médicos da rede municipal de saúde.

A rede de serviços do Sistema Único de Saúde (SUS) de Curitiba atende uma população de 1.908.359 habitantes, segundo dados do IBGE de 2017 (SAÚDE, 2018). Conta com 111 Unidades de Saúde, sendo 44 Unidades Básicas de Saúde (UBS), 67 Unidades de Saúde com Estratégia da Saúde da Família (US/ESF), nove Unidades de Pronto Atendimento, 12 Centros de Atenção Psicossocial (CAPS), cinco Unidades Especializadas/Especialidades Médicas, três Centros de Especialidades Odontológicas, dois Hospitais, um Laboratório de Análises Clínicas, uma Central de Vacinas, cinco Residências Terapêuticas, um Centro de Zoonoses e 68 Espaços Saúde.

Dados publicados em formato aberto permitem que qualquer cidadão desenvolva aplicações ou visualizações que facilitem a análise destes dados (SANTOS, 2016). Tais aplicações podem promover a melhoria de serviços públicos, o que contribuiria para uma maior participação da sociedade junto ao governo.

É possível aplicar nesse tipo de dado um processo de descoberta de conhecimento ou *Knowledge-Discovery in Databases* (KDD) e com isso identificar informações úteis.

O processo de extração de conhecimento de bases de dados busca encontrar conhecimento a partir de um conjunto de dados, para que ele possa ser utilizado em um processo decisório. Um requisito importante é que o conhecimento descoberto seja útil e de interesse para os usuários e que seja apresentado de forma compreensível, que apoie a tomada de decisões (FAYYAD et al., 1996a).

No Portal de Dados Abertos da PMC, na seção Saúde, são disponibilizadas mensalmente quatro bases de dados para serem descarregadas.Estas bases contêm os dados referentes ao perfil de atendimento dos profissionais que trabalham nas Unidades Básicas de Saúde, nas Unidades de Pronto Atendimento e de todos os outros equipamentos especializados da rede municipal de saúde (DADOS, 2018).

Por conterem dados de toda a rede de atendimento, os arquivos das bases são grandes, aproximadamente 600 GB, pois possuem centenas de milhares de registros. Essa quantidade de dados faz necessário o uso de técnicas específicas para análise e processamento dos dados, bem como de técnicas de interpretação destes dados.

Nesse trabalho, foi realizado um estudo no qual o processo de descoberta do conhecimento foi aplicado aos dados extraídos do E-Saúde. Foram explorados os dados relativos à resolutividade no atendimento ao usuário da rede de saúde, mais especificamente no que se refere as solicitações de exames em uma consulta.

Como contribuição, este trabalho estuda a viabilidade, pela análise da aplicação de técnicas de mineração de dados na base de dados pública de saúde disponível no portal de dados abertos na PMC, para obtenção de informações que sejam úteis e relevantes para os gestores do Sistema de Saúde.

1.1. O PROBLEMA

O problema ao qual este trabalho se dedicou foi o de estudar a base de dados E-Saúde, sistema de informática da Saúde de Curitiba, no esforço de descobrir, se esta base é capaz de fornecer conhecimentos, que auxiliem na tomada de decisão dos gestores das Unidades de Saúde.

1.2. OBJETIVO GERAL

O objetivo geral deste trabalho foi o de analisar a possibilidade de explorar a base de dados do E Saúde, usando mineração de dados, para descobrir conhecimento potencialmente útil para a tomada de decisões dos gestores das Unidades de Saúde do sistema de Saúde da cidade de Curitiba.

1.3. OBJETIVOS ESPECÍFICOS

Para atingir o objetivo geral, são considerados os seguintes objetivos específicos:

- Importar, pelo uso de aplicação web especificamente desenvolvida, os dados dos arquivos disponibilizados no Portal de Dados Abertos para um banco de dados MySQL.
- Criar, pelo uso de aplicação web especificamente desenvolvida, os arquivos .ARFF para serem utilizados no software Weka.
- Criar atributos novos com o intuito de auxiliar as análises dos dados.
- Realizar experimentos para testar o desempenho, em relação a acurácia, dos algoritmos de classificação aplicados aos registros da base de dados.

1.4. ESTRUTURA DO TEXTO

Este trabalho está dividido em cinco capítulos. No Capitulo 1 está a introdução no qual são apresentados o problema, o objetivo geral e os objetivos específicos dessa pesquisa.

O Capítulo 2 contém a revisão bibliográfica dos tópicos necessários para realização da pesquisa.

O Capítulo 3 apresenta os métodos para a preparação dos dados, a criação de novos atributos e a seleção dos melhores atributos para a classificação. Também apresenta a implementação dos métodos propostos.

No Capítulo 4 define o escopo do trabalho, descrevendo os objetivos dos experimentos de mineração de dados realizados nos dados referentes às unidades de saúde de dois distritos sanitários de Curitiba. Ainda neste capítulo são apresentadas as análises dos resultados obtidos a partir dos experimentos realizados.

Por fim, no Capítulo 5 são apresentadas as considerações finais e as conclusões sobre as contribuições deste trabalho, bem como suas limitações. E ainda, são identificadas possibilidades de continuidade do trabalho.

2. REFERENCIAL TEÓRICO

Considerando a saúde como um direito de todos e um dever do Estado, é função do Sistema Único de Saúde, SUS, dispor de condições para a promoção e a recuperação da saúde do indivíduo e da comunidade, respeitando os princípios da universalidade, integralidade e resolutividade (SAUDE, 2018).

O SUS está estruturado em três níveis hierárquicos de atenção à saúde que se complementam: primário, médio e de alta complexidade. Cada nível apresenta limites com relação à complexidade e à capacidade de resolução dos problemas que pode tratar (OLIVEIRA et al., 2015).

Os entes que representam cada um destes níveis são, respectivamente, as Unidades Básicas de Saúde (UBS), as Unidades de Pronto Atendimento (UPA) e os Hospitais.

2.1. UNIDADES BÁSICAS DE SAÚDE (UBS)

As Unidades Básicas de Saúde estão próximo de onde as pessoas moram e/ou trabalham, desempenhando assim um papel central na garantia do acesso da população a atenção à saúde. Nelas são realizados exames, consultas e acompanhamento médico, bem como a entrega de remédios e a aplicação de vacinas. Por isso, as UBS são consideradas portas de entrada do usuário no SUS (ALVES FILHO e BORGES, 2014).

Quando necessário, na UBS é feito o encaminhamento do paciente a outros serviços de saúde. Em Curitiba, as UBS começam a funcionar às 07h00 e podem operar com horários de fechamento variando de 17h00 a 22h00, de segunda à sexta-feira (ATENÇÃO, 2018).

A Atenção Primária em Curitiba possui dois modelos de Unidades de Saúde: as Unidades Básicas de Saúde (UBS) e as Unidades de Saúde da Família (ESF) (NUNES et al., 2012). Ambas abrigam equipes de profissionais compostas por: agente comunitário de saúde, auxiliar de enfermagem, auxiliar de saúde bucal, cirurgião dentista, médico, enfermeiro, farmacêutico, fisioterapeuta, psicólogo e outros profissionais tanto de nível superior quanto de técnico.

Uma ESF busca promover a qualidade de vida com a assistência provida por médico especialista treinado para atender a população em a todas as fases do ciclo de vida, desde o nascimento, passando pela gravidez até à velhice (ESF, 2018). A USF busca intervir nos fatores que colocam a saúde em risco, como falta de atividade física, má alimentação e uso de tabaco.

Já na UBS, os médicos atendem a população de acordo com a sua especialidade que pode ser Pediatria, Clínica Geral, Ginecologia/Obstetrícia, Enfermagem e Odontologia.

2.2. UNIDADES DE PRONTO ATENDIMENTO (UPA)

As UPA estão inseridas dentro da Rede de Atenção para atender as urgências e as emergências. Elas têm por objetivo concentrar os atendimentos de saúde de complexidade intermediária, trabalhando em conjunto com a atenção básica, as UBS e a atenção hospitalar (UPA, 2018).

A UPA deve proporcionar à população um atendimento à saúde capaz de diminuir as filas nos prontos socorros de hospitais. Nela, é a gravidade do risco e não a ordem de chegada que determina a rapidez com que o paciente será atendido. As UPA funcionam 24 horas, todos os dias da semana (UPA, 2018).

Neste trabalho não foram analisados os dados de outros entes da rede de atenção, como os hospitais, os Centros de Especialidades, os Centros de Apoio Psicossocial ou os Centro de Especialidades Odontológicas, entre outros.

2.3. A REDE DE ATENÇÃO DE CURITIBA

A Rede de Atenção de Curitiba atende tanto o usuário da cidade quanto de outros municípios da Região Metropolitana. A rede conta com 141 serviços próprios, dentre os quais estão 109 Unidades Básicas de Saúde (UBS), sendo que 65 atuam no modelo Saúde da Família e nove como Unidades de Pronto Atendimento (UPA) (Figura 1).

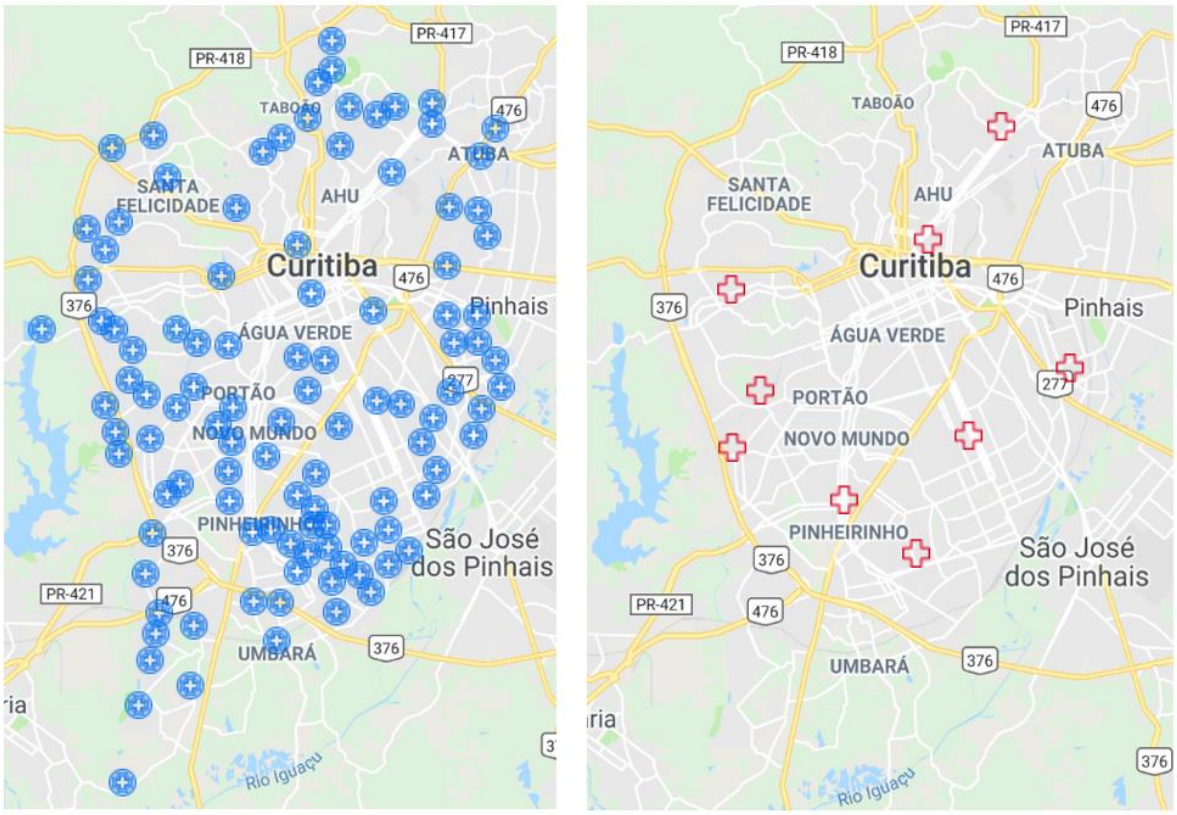


Figura 1 - Distribuição das UBS (esquerda) e UPA (direita) em Curitiba.
Fonte: Autoria própria.

2.4. DISTRITOS SANITÁRIOS (DS)

O Distrito Sanitário (DS) é uma área geográfica que comporta uma população com características epidemiológicas e sociais semelhantes. No DS é encontrado um agrupamento de várias unidades, podendo incluir uma UPA, e outros equipamentos (CARDOSO et al., 2013).

Algumas das atribuições do Distrito Sanitário são a execução das ações e serviços de saúde, o planejamento de políticas e outros mecanismos que buscam ampliar o acesso aos serviços de assistência à saúde e a otimização dos recursos.

A Rede de Saúde de Curitiba está organizada em 10 DS. Na tabela 1 são apresentados os distritos, bem como o número de UBS, de UPS e de outros equipamentos que compõem cada distrito.

Tabela 1 - Distritos Sanitários.

Distrito Sanitário	UBS	UPA	Outros
Bairro Novo	12	1	2
Boqueirão	14	1	2
Boa Vista	19	1	3
Cajuru	12	1	2
CIC	17	1	2
Matriz	3	0	5
Pinheirinho	11	1	3
Portão	6	1	4
Santa Felicidade	9	1	2
Tatuquara	8	0	0

Fonte: (DISTRITOS, 2018)

As UBS e as UPA dos Distritos Sanitários do Boqueirão e do Cajuru, os quais tiveram dados utilizados neste trabalho, estão distribuídas por nove bairros de Curitiba, sendo quatro do DS do Boqueirão: Boqueirão, Alto Boqueirão, Xaxim e Hauer e cinco do DS do Cajuru: Capão da Imbuia, Cajuru, Guabirotuba, Jardim das Américas e Uberaba (Tabela 2 e Figura 2).

Tabela 2 - Distribuição das UBS e UPA dos Distritos Sanitários nos bairros.

	Unidades DS Boqueirão	Bairro	Unidades DS Cajuru	Bairro
1	US Érico Veríssimo	Alto Boqueirão	US Alvorada	Uberaba
2	US Esmeralda	Xaxim	US Cajuru	Cajuru
3	US Eucaliptos	Alto Boqueirão	US Camargo	Cajuru
4	US Irmã Teresa Araújo	Boqueirão	US Iracema	Capão da Imbuia
5	US Jardim Paranaense	Alto Boqueirão	US Lotiguaçu	Uberaba
6	US Menonitas	Xaxim	US Salgado Filho	Uberaba
7	US Moradias Belém	Boqueirão	US São Domingos	Cajuru
8	US Pantanal	Alto Boqueirão	US São Paulo	Uberaba
9	US São Pedro	Xaxim	US Solitude	Cajuru
10	US Tapajós	Xaxim	US Trindade	Cajuru
11	US Vila Hauer	Hauer	US Trindade II	Cajuru
12	US Visitação	Boqueirão	US Uberaba de Cima	Uberaba
13	US Waldemar Monastier	Boqueirão	UPA Cajuru	Cajuru
14	US Xaxim	Xaxim		
15	UPA Boqueirão	Boqueirão		

Fonte: (DISTRITOS, 2018)

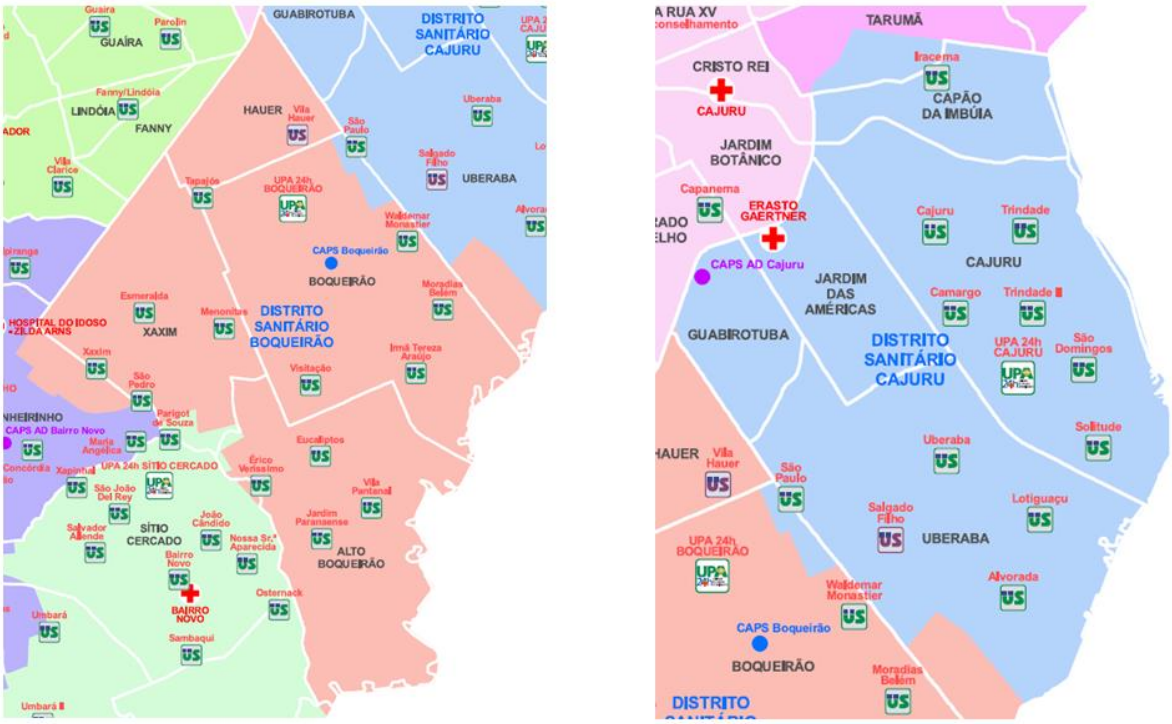


Figura 2 - Distribuição nos Distritos Sanitários do Boqueirão e Cajuru.
Fonte: (CARTEIRA, 2018).

2.5. DADOS ABERTOS

A transparência governamental é um dos fundamentos da democracia e uma das formas pelas quais ela se materializa é pelo acesso do cidadão às informações do governo (SILVA et al., 2014). A política de abertura de dados pode potencialmente melhorar a gestão pública, incentivar a participação social e estimular a inovação.

A Instrução Normativa nº 4 de 12 de Abril de 2012 instituiu a política dos dados públicos abertos no Brasil. Nela, o dado público é todo aquele gerado ou sob a guarda do governo que não seja restrito por qualquer legislação. Já o dado aberto é um dado público digitalizado e acessível na web (INDA, 2018).

Todo dado público pode tornar-se um dado aberto (PORTAL, 2018). Considerando que quase todos os dados governamentais são públicos, existe um campo amplo a ser explorado por políticas e práticas visando à disponibilização de dados na forma aberta para a população.

De acordo com a *Open Knowledge*, uma organização sem fins lucrativos que promove conhecimento livre, os dados são abertos quando qualquer pessoa pode usá-los de forma livre, reutilizá-los e redistribuí-los da forma que desejar. Dados abertos podem estar sujeitos, no máximo, à exigência de creditar a sua fonte, sem restrição de licenças, patentes ou mecanismos de controle (OKF, 2018).

O governo federal mantém o Portal Brasileiro de Dados Abertos (DADOS ABERTOS, 2018) com cerca de 5.427 bases de dados disponíveis, sendo 136 referentes ao Ministério da Saúde (AVENTURIER e ALENCAR, 2016).

A disponibilização dos dados abertos implica que os dados estarão contextualizados e conexos. Segundo o grupo de trabalho denominado *Open Government Working Group*, existe um conjunto de características que permitem que um dado seja considerado um dado aberto governamental (OPEN GOVERNMENT, 2018). São eles:

1. Completos.
2. Primários.
3. Atuais.
4. Acessíveis.
5. Processáveis por máquina.
6. Acesso não discriminatório.
7. Formatos não proprietários.
8. Livres de licenças.

2.5.1. Portal de Dados Abertos de Curitiba e o E-Saúde

O Portal de Dados Abertos de Curitiba tem o objetivo de prover a busca e o acesso a todos dados públicos abertos de Curitiba (PORTAL, 21018). O E-Saúde é o sistema de informática utilizado pela SMS de Curitiba, que pode ser acessado via Portal de Dados Abertos de Curitiba. O E-Saúde foi desenvolvido e é mantido pelo ICI, Instituto das Cidades Inteligentes (E-SAÚDE, 2018).

O E-Saúde foi desenvolvido para atender as diretrizes do Ministério da Saúde e permite o gerenciamento dos recursos municipais relacionados à área da saúde. O sistema provê apoio aos procedimentos prestados pela SMS de Curitiba em sua Rede de Atenção. Tais procedimentos englobam desde procedimentos da atenção

básica, até os mais especializados, tanto ambulatoriais e laboratoriais quanto hospitalares.

No Portal de Dados Abertos de Curitiba estão disponíveis quatro bases de dados com informações extraídas do E-Saúde, que foram objetos do presente estudo.

Para o restante do trabalho, as quatro bases serão tratadas como se fossem uma única base de dados (a base de dados E-Saúde), apesar de cada uma delas conter um subconjunto dos dados gerados pelo sistema como um todo. Na figura 3, é mostrada uma tela do sistema E-Saúde.

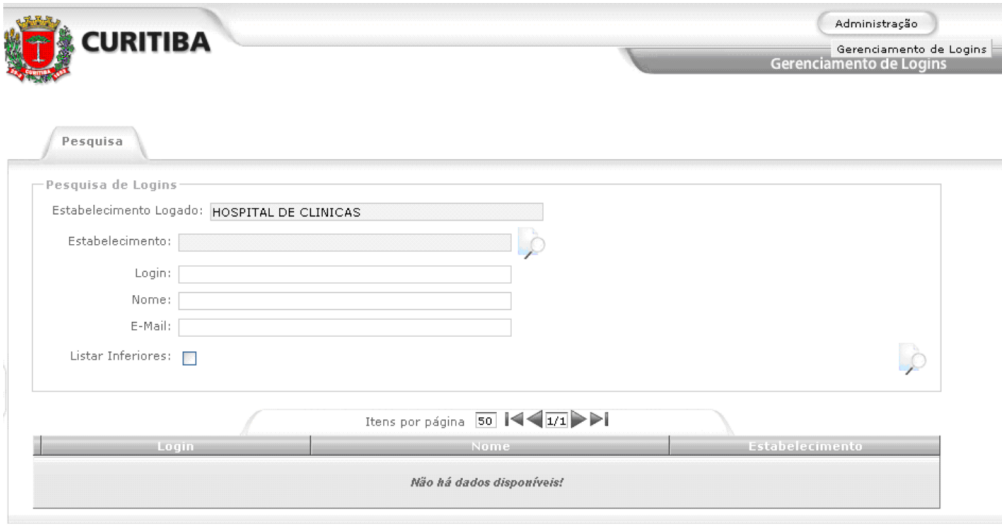


Figura 3 - Tela do sistema de informática E-Saúde.

Fonte: Manual do E-Saúde.

2.6. DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

A crescente difusão das tecnologias digitais cria dados de forma acelerada, fazendo com que haja também o acelerado crescimento na quantidade de dados armazenados (GABARDO, 2015).

Tal quantidade de dados pode ser analisada para a identificação de informações úteis que podem, por sua vez, levar a uma melhor tomada de decisões. No entanto, para extrair informação e conhecimento dessa quantidade de dados, é necessário que se use métodos, técnicas e ferramentas que automatizem a extração de informações e conhecimento desses dados.

A área de conhecimento que trata desses métodos, técnicas e ferramentas é a Descoberta de Conhecimento em Bases de Dados, ou *Knowledge Discovery in Databases* (KDD) (SILVA, 2007).

O KDD permite extrair conhecimento de grandes bases de dados especializadas em diferentes áreas (SOARES JUNIOR e QUINTELLA, 2005). Dentro do KDD, o papel da mineração de dados é o da identificação de padrões de dados válidos, não-triviais, potencialmente úteis e compreensíveis contidos em uma base de dados (FAYYAD et al., 1996a).

A mineração de dados consegue detectar informações implícitas armazenadas nas bases de dados, transformando-as em conhecimento, enquanto os métodos convencionais tratam apenas as informações explícitas.

Ainda, o KDD pode ser definido como multidisciplinar, pois inclui áreas do conhecimento como aprendizado de máquina (algoritmos de busca, técnicas de modelagem, teorias de aprendizagem), estatística (ideias de amostragem e teste de hipóteses) e banco de dados (armazenamento, indexação e processamento de consultas) (Figura 4) (COSTA et al., 2018).

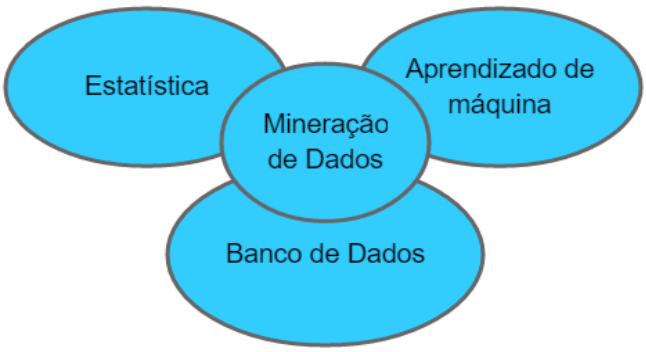


Figura 4 - Disciplinas do KDD

Fonte: Figura adaptada de (COSTA et al., 2018).

O KDD é um processo que envolve várias fases, com iterações entre elas. O fluxo básico do KDD está ilustrado na Figura 5.

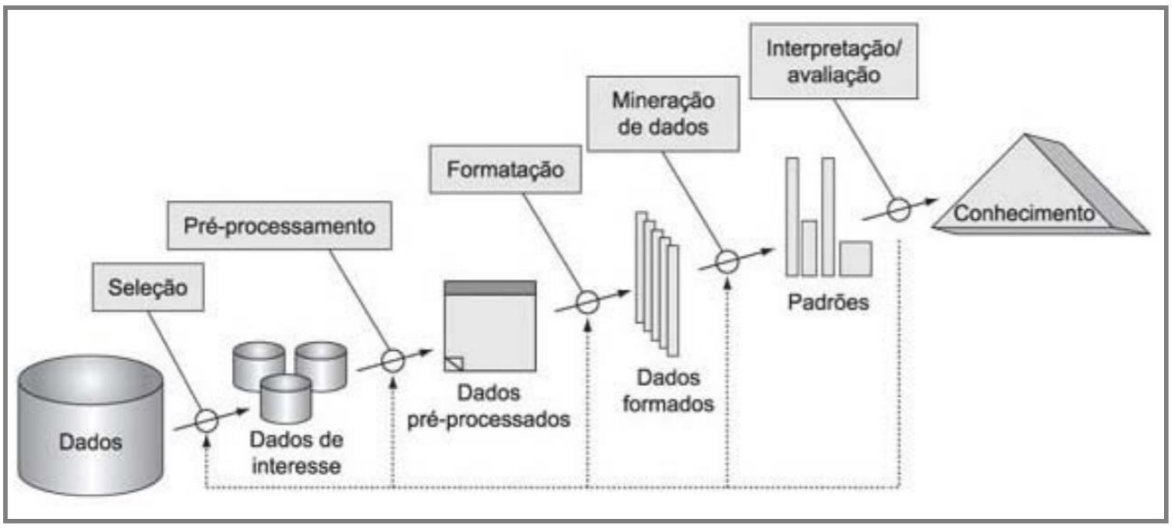


Figura 5 - Fases do Processo do KDD.
Fonte: (FAYYAD et al., 1996a).

O processo de KDD é composto por fases sequenciais, mas é também iterativo. Pode haver retorno às fases anteriores e às descobertas já feitas. Assim, o processo pode levar a novas hipóteses e descobertas. O KDD pode ser divido em seis fases (Quadro 1) (ELMASRI e NAVATHE, 2005).

Quadro 1 - Fases do KDD.

Fase	Ação
1 - Integração dos dados	Combinação de diferentes fontes de dados, quando for o caso
2 - Seleção dos dados	Recuperação de dados relevantes para a análise do banco de dados
3 – Pré-processamento dos dados	Remoção dos ruídos e dados irrelevantes
4 - Formatação dos dados	Consolidação no formato apropriado para mineração
5 - Mineração de dados	Aplicação de algoritmos dedicados a extrair padrões dos dados
6 - Avaliação e representação do conhecimento	Emprego de técnicas de visualização e representação do conhecimento extraído para o usuário

Fonte: (ELMASRI e NAVATHE, 2005).

A literatura identifica como objetivo principal do KDD, extrair conhecimento de bases dados. A descoberta do conhecimento envolve uma sequência de fases a serem seguidas para se alcançar o objetivo. Embora as fases estejam encadeadas

de forma linear, em muitos casos, ao se concluir uma fase, em vez de se seguir para a próxima, voltar para a anterior, ou até mesmo, fases anteriores, no intuito de se refinar o processo.

Também é comum algumas fases acontecerem de forma simultânea, como, por exemplo, a seleção e pré-processamento. Ou nem mesmo acontecem, por exemplo, quando se tem apenas uma base de dados, a fase da Integração não se faz necessária.

Assim, é comum utilizar-se abordagens mais concisas do KDD, orientada a um objetivo (SHIBA, 2008). O processo pode ser agrupado em três etapas principais, conforme visto no quadro 2.

Quadro 2 - Etapas do KDD deste trabalho.

Fase	**Etapa**
Pré-processamento	Compreensão do domínio, escolha e seleção de dados, limpeza e preparação (transformação) dos dados
Mineração de dados	Escolha das tarefas de a mineração, escolha dos algoritmos apropriados, indução de modelos e testes
Pós-processamento	Visualização e interpretação dos resultados, avaliação dos resultados

Fonte: (SHIBA, 2008).

2.6.1. Terminologia das Bases de Dados

Bases de dados são coleções eletrônicas que armazenam grandes quantidades de informação. São organizadas de forma estruturada possibilitando a consulta rápida e facilitada dos dados nelas armazenados. Dentro das bases de dados, as informações podem ser armazenadas de várias formas como, por exemplo, em tabelas.

Fisicamente, as tabelas são compostas por linhas e colunas, cujos cruzamentos constituem os campos. Quando se consulta uma tabela, as informações são retornadas em blocos, com tamanhos variáveis, de acordo com o alcance da consulta realizada. É usual chamar uma linha desse bloco retornado de tupla ou registro, sendo que cada registro é uma instância da tabela (ou entidade). As colunas, nas quais realmente estão os dados, podem ser chamadas de atributos (DATE, 2004).

2.6.2. Mineração de Dados (MD)

Mineração de Dados é uma fase do KDD na qual se aplicam algoritmos para descobrir padrões e extrair modelos existentes nos dados (FAYYAD et al., 1996b). Pela aplicação dos algoritmos, é possível encontrar relações de similaridade ou discordância entre dados, transformando assim informações aparentemente ocultas, em explícitas para a tomada de decisão ou avaliação de resultados (VIEIRA,2015).

Além de produzir conhecimento, que pode ser utilizado nas organizações para definição de suas estratégias e para o processo de tomada de decisões, a MD também pode produzir modelos de previsão ou modelos preditivos. Tais modelos marcam ou classificam novos exemplos de acordo com os padrões existentes nos atributos que foram utilizados durante a criação destes modelos (KHABAZA, 2010).

2.6.3. Dados Médicos

A MD pode ser aplicada em diversas áreas do conhecimento, incluindo a saúde, que por sua vez, é a área deste estudo. Contudo, aplicar técnicas MD a dados médicos é um processo desafiador (FERREIRA, 2015). Por exemplo, muitas vezes operadores preenchem os sistemas de informática de forma incompleta ou falha, tendendo a produzir bases de dados esparsas, com dados altamente variados, demandando o uso de diferentes técnicas e ferramentas para que as bases possam ser exploradas de maneira que produzam algum resultado eficaz.

Existem ainda restrições éticas, legais e sociais relativas à privacidade e à validação clínica dos achados. Assim, na maioria das vezes, os dados que efetivamente chegam para análise são um subconjunto dos dados originais.

Na área de saúde, guardam-se informações relevantes dos pacientes, dos diagnósticos, dos tratamentos, dos exames, dos laudos e dos medicamentos (COSTA, 2007). Portanto, a quantidade de dados primários disponível é tal que torna o processo de encontrar informação implícita manualmente praticamente impossível. Dentro desse contexto, a mineração de dados pode ser usada para contribuir para a tomada de decisão de médicos e enfermeiros.

2.7. TAREFAS DE MINERAÇÃO

A MD engloba diversas técnicas, cada uma com suas vantagens e desvantagens, que podem ser usadas para tratar problemas de diferentes naturezas (COSTA et al., 2018). Um prévio conhecimento de cada técnica orienta a escolha de uma ou mais delas para tratar o problema em questão.

Dependendo do problema a ser tratado, as duas principais metas que podem ser alcançadas usando MD são a Previsão e a Descrição (FAYYAD et al., 1996a) (Figura 5).

A Previsão usa atributos existentes na base de dados para prever valores desconhecidos ou futuros de outros atributos que se tenha interesse. Já a Descrição foca na busca de padrões interpretáveis por seres humanos, descrevendo as relações existentes entre os atributos (HAN et al., 2011).

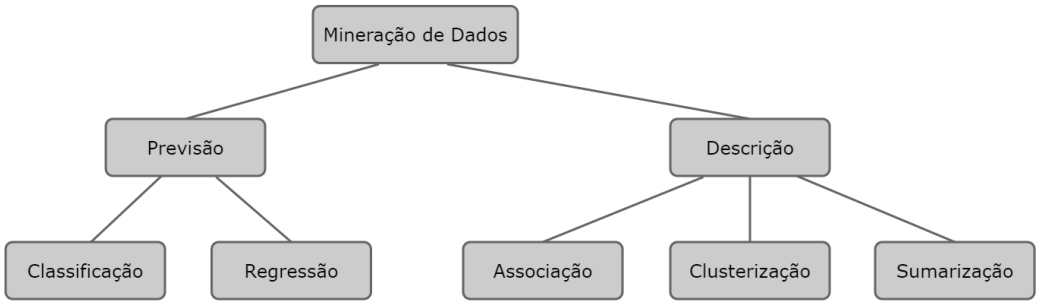


Figura 6 - Tarefas da Mineração de Dados.
Fonte: Figura adaptada de (HAN et al., 2011).

No contexto da MD, uma tarefa consiste em uma técnica para a Descoberta de um Conhecimento que possui suas próprias características. Devido a isso, é necessário um conhecimento prévio de como cada tarefa funciona e como ela pode ser mais bem aplicada. As tarefas podem ser de (SILVA, 2007):

• **Classificação:** objetiva encontrar modelos que evidenciem classes a partir dos atributos apresentados, de forma que possam ser aplicados a atributos não classificados, identificando possíveis tendências.

• **Regressão**: também chamada de Estimativa, é usada quando o atributo a ser classificado apresenta valores numéricos (contínuos). Os vários atributos de uma instância combinados por uma relação definem o valor de atributo definido como saída.

• **Associação:** consiste em identificar quais atributos estão relacionados, ou seja, que tendem a ocorrer ao mesmo tempo durante uma mesma transação, criando regras, nas quais a existência de um ou mais atributos implica em outro.

• **Clusterização:** também chamada de Agrupamento, consiste em agrupar instâncias com atributos similares de uma população maior em subconjuntos, de modo que cada subconjunto seja formado por instâncias com atributos com características semelhantes.

• **Sumarização:** consiste em descrever, de uma maneira compacta, um determinado conjunto de instâncias, apresentando as suas principais características.

Em virtude das características dos dados armazenados na base de dados E-Saúde, neste trabalho foi utilizada a tarefa de classificação.

2.8. APRENDIZADO DE MÁQUINA

O aprendizado indutivo é efetuado a partir de raciocínio sobre exemplos fornecidos por processo externo ao sistema de aprendizado. Pode ser dividido em supervisionado e não-supervisionado (REZENDE, 2005).

No aprendizado supervisionado, é fornecido um algoritmo de aprendizado, ou indutor e um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido.

Em geral, cada exemplo é descrito por um vetor de valores de características, ou atributos, e pelo rótulo da classe associada. O objetivo do algoritmo de indução é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados.

Existem vários paradigmas de Aprendizado de Máquina (AM), tais como: Simbólico, Estatístico, Baseado em Exemplos, Conexionista e Evolutivo. Os sistemas de aprendizado simbólico (MONARD e PRATI, 2005), buscam aprender construindo representações simbólicas de um conceito pela análise de exemplos e contraexemplos deste conceito. As representações simbólicas estão tipicamente na forma de alguma expressão lógica, Árvore de Decisão ou Regras.

2.8.1. Classificadores

A classificação é uma das tarefas mais utilizadas na MD, por ser uma das tarefas cognitivas mais utilizadas na compreensão e percepção do ambiente em que se opera (AMARAL, 2016).

A tarefa de classificação pode ser entendida como o processo sistemático de organizar itens de um determinado grupo de entrada, separando-os em categorias pré-definidas usando analogias e características comuns (LIBRELOTTO e MOZZAQUATRO, 2013),

O principal objetivo da classificação é que seja construído um classificador que represente a descoberta de algum tipo de relação entre as variáveis disponíveis no conjunto de dados de entrada. Nessa tarefa, deve ser identificada, entre as variáveis de entrada, aquela que representa a classe a ser predita pelo classificador, a qual se denomina “variável classe” (FREITAS e LAVINGTON, 1998).

Os dados de entrada da tarefa de classificação são um conjunto de registros. Cada registro, também conhecido como instância, é caracterizado por uma dupla (*x,y*), onde *x* é o conjunto de atributos e *y* o atributo especial, designado rótulo de classe (também chamado de atributo alvo ou de categorização). Os atributos podem ser de qualquer tipo, contudo o rótulo de classe deve ser um atributo discreto (REZENDE, 2005).

Assim, a classificação é a tarefa de aprender uma função alvo *f(x)*, que mapeie cada conjunto de atributos *x* para um dos rótulos de classe *y* pré-determinados. A função alvo também é conhecida informalmente como *modelo de classificação*.

Uma técnica de classificação é uma abordagem sistemática para a construção de modelos de classificação a partir de um conjunto de dados de entrada. Exemplos incluem classificadores de Árvores de Decisão, baseados em regras, redes neurais, entre outros.

Cada técnica emprega um algoritmo de aprendizagem para identificar um modelo que seja mais apropriado para o relacionamento entre o conjunto de atributos e o rótulo de classe dos dados de entrada (TAN et al., 2009).

Um classificador extraído de um conjunto de dados serve a dois propósitos: predição de um valor e entendimento da relação existente entre aquela indicada

como a classe e as demais variáveis do conjunto disponível (BREIMAN et al.,1984). Como exemplo, pode-se ter uma aplicação na área da saúde, na qual um médico poderia classificar alguns de seus pacientes em duas classes: “tem” ou “não tem” determinado diagnóstico.

Para cumprir o segundo propósito, é exigido do classificador que ele não apenas classifique, mas, também, explicite as relações entre os atributos, baseado em conhecimento extraído da base de dados, de forma compreensível (CARVALHO et al., 2012). Assim, pode-se aplicar o modelo obtido para predizer a classe de um atributo referente a registros que ainda não foram classificados.

2.8.2. Treinamento e Teste

A indução de um modelo usando todo o conjunto de dados de entrada ocasiona um problema chamado sobre ajuste (*overfitting*). Este problema ocorre quando o modelo está excessivamente ajustado ao conjunto de treinamento (HAN et al., 2011).

O sobre ajuste representa um problema, pois quando acontece, o modelo funciona bem para os dados existentes, mas não consegue generalizar para novas situações.

O sobre ajuste pode fazer com que o modelo induzido não seja capaz de generalizar o caso tratado, fazendo com que novos dados não sejam classificados corretamente quando analisados (OLIVEIRA JÚNIOR, 2015). O modelo induzido por sobre ajuste não foi treinado de forma mais abrangente, o que o restringe apenas ao conjunto original de dados. De forma mais exata, usar o conjunto de dados inteiro produz uma estimativa muito otimista do classificador utilizado, com resultados superiores a realidade.

Para evitar ou diminuir o sobre ajuste, é feita uma separação no conjunto de dados de entrada em dois subconjuntos, um para treinamento e outro para validação (teste) do modelo induzido (WITTEN et al., 2011).

O processo de treinamento usa o subconjunto separado para o treino, no qual estão presentes todos os atributos de cada ocorrência (instância), inclusive o atributo que se deseja classificar como saída, para construção de um modelo (TAN et al., 2009). Já para realizar o teste do modelo obtido, usa-se o outro subconjunto.

Nele é aplicado o modelo construído com o intuito de confirmar a consistência da predição obtida com os dados de teste em relação a saída esperada para o modelo.

Para dividir um conjunto de dados de entrada em dois subconjuntos, um para treinamento e outro para teste, existem vários métodos. Um primeiro método consiste em se utilizar uma parte do conjunto de dados original para um pré-processamento inicial e, depois já com um modelo estabelecido, utilizar esse subconjunto para realizar o teste no subconjunto que sobrou do conjunto original.

Outro método é conhecido como *Percentage Splits*. Ele consiste em dividir o conjunto total de dados em dois subconjuntos, o maior deve ser usado para treinamento. Esse subconjunto é usado durante a fase de pré-processamento com o intuito de criar um modelo otimizado em termos de resultados estatísticos. O outro subconjunto, o menor, é usado para os testes (SILVA et al., 2016).

Um terceiro método é a validação cruzada (*cross-validation*). Esse foi o método utilizado neste trabalho. Ele consiste em dividir o conjunto total de dados em diferentes subconjuntos de igual tamanho, sendo cada um deles mutuamente exclusivo (PEREIRA e SOUZA JUNIOR, 2017).

A escolha de quais técnicas (e algoritmos) de MD aplicar depende da tarefa de mineração a ser realizada. As exigências das tarefas de mineração e as suas características influenciam a viabilidade da sua aplicação em um determinado problema.

A tarefa de classificação, utilizada neste trabalho, conta com duas técnicas de classificação. Em função das características de saída apresentadas e dos modelos mais recorrentes no contexto de mineração de dados médicos, foram usados um classificador de Árvore de Decisão e um baseado em Regras.

2.8.3. Árvores de Decisão

Uma das técnicas de classificação mais utilizadas na MD é a das Árvores de Decisão (AD). A técnica das AD apresenta vantagens como: ser de construção simples e rápida, ser adequada a problemas com muitas dimensões e por ser de fácil representação e visualização (COSTA, 2007). As AD são classificadores por concepção, pois utilizam a abordagem de subdividir para classificar.

Algoritmos que induzem Árvores de Decisão pertencem à família de algoritmos *Top Down Induction of Decision Tree* - TDIDT. Uma AD é uma estrutura de dados definida recursivamente como:

- Um nó folha que corresponde a uma classe ou
- Um nó de decisão do teste que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma sub árvore.

Desta forma, uma AD é uma representação de um conjunto de regras de classificação por meio de nós. Parte-se de um nó inicial até se chegar aos nós terminais, que representam a classe de uma determinada instância (ROMERO et al., 2008).

A Figura 7 apresenta o exemplo de uma AD para um diagnóstico de um paciente. Cada elipse é um teste de um atributo para um determinado conjunto de dados de pacientes. Cada retângulo representa uma classe, ou seja, um diagnóstico. Para diagnosticar (classificar) um paciente, começa-se pela raiz, seguindo cada teste até que uma folha seja alcançada.

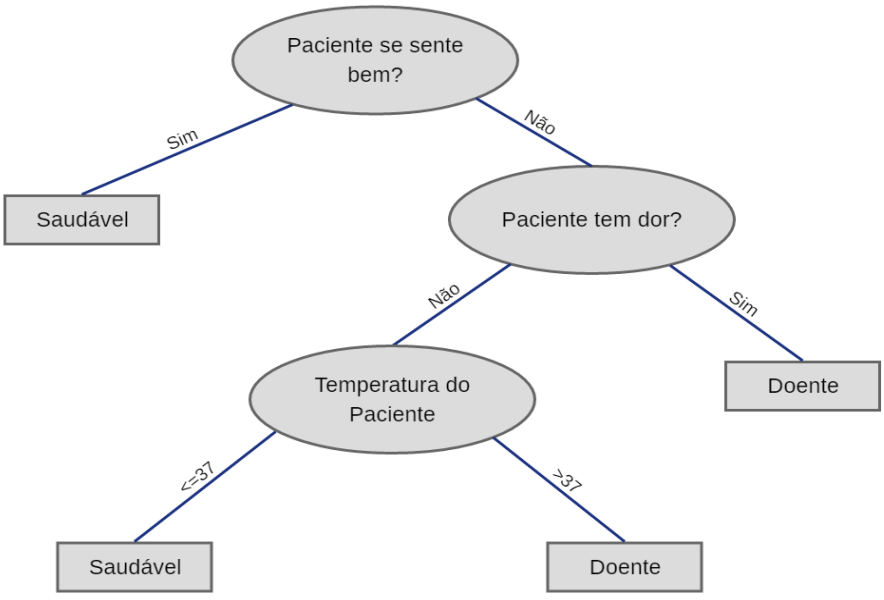


Figura 7 - Exemplo de Árvore de Decisão.

Fonte: (REZENDE, 2005).

Basicamente, uma AD é um algoritmo recursivo de busca gulosa, que procura, sobre um conjunto de atributos, aqueles que “melhor” dividem o conjunto dos dados

em subconjuntos. Inicialmente, todos os dados são colocados em um único nó, chamado de raiz.

A seguir, um atributo preditivo é escolhido para representar o teste desse nó e, assim, dividir os exemplos em subconjuntos de exemplos. Esse processo se repete recursivamente até que todos os exemplos já estejam classificados ou então até que todos os atributos preditivos já tenham sido utilizados.

As AD são consideradas modelos de fácil compreensão porque é possível acompanhar o processo de raciocínio que leva a cada conclusão. Elas podem funcionar com atributos numéricos ou nominais.

Uma AD é considerada ideal quando representa o maior número de dados com o menor número de níveis ou perguntas (HAN et al., 2011). Quando o treinamento da AD está finalizado, é possível alimentá-la com novos casos para que eles sejam classificados.

A AD foi utilizada neste trabalho porque é um método simbólico que representa por meio de expressões o que é aprendido sobre os atributos, identificando aqueles mais fortemente relacionados ao atributo a ser classificado (NAKAMURA et al., 2016).

2.8.4. Classificadores Baseados em Regras

A obtenção de Regras de Classificação é amplamente utilizada como forma de extração de conhecimento para indução de modelos de aprendizagem (SHIBA, 2008). As RC são consideradas um método simbólico que representa por meio de expressões o que é aprendido sobre os atributos de entrada (NAKAMURA et al., 2016). As regras são usadas para representar informação ou pedaços de conhecimento (HAN et al., 2011).

Em um classificador baseado em Regras, o modelo aprendido é exibido na forma de um conjunto de regras do tipo “SE-ENTÃO”. Por exemplo, “SE as condições 1, 2 e 3 forem cumpridas, ENTÃO o desfecho x será o resultado, com uma confiança de y”. Desta forma, SE os valores assumidos atendem as condições do antecedente da regra, ENTÃO recebe a classe indicada pelo valor do atributo da classificação. A representação de uma regra possui o seguinte formato:

Se <*Condição(ções)*> **Então** <*Classificação*>

A *Condição* é constituída por uma ou mais expressões condicionais envolvendo os atributos existentes no banco de dados com seus respectivos valores.

Já a *Classificação* recebe um dos valores do atributo a ser classificado. Assim, a tarefa de classificação induz (cria) um modelo capaz de fazer avaliações futuras do atributo a ser classificado (SILVA, 2007). Cada regra obtida pode ser expressa da seguinte maneira:

r_i: (Condição) -> y_i

Onde o lado esquerdo da regra é chamado de antecedente ou pré-condição. Já o lado direito da regra é o consequente da regra, que contém a classe y_i prevista.

A tabela 3 apresenta um conjunto de Regras geradas com fragmentos de atributos da base de dados E-Saúde deste trabalho.

Tabela 3 - Regras geradas com atributos da base E-Saúde.

	Regra	Ocorrências Total/Erros
1	**Se** FaixaIdade=18-35 **Então** Tipo Médico=ClínicoGeral	864/133
2	**Se** Sexo=Feminino **E** Idade > 60 **Então** TipoMedico=Geriatra	694/125
3	**Se** TipoCasa=Simples **E** Transporte=ônibus **Então** EquipamentoSaude=Unidade de Saúde	399/201
4	**Se** CBO=MedicoDaFamilia **E** IdadeFaixa=Criança **E** Encaminhamento=Ginecologista **Então** Exames=Sim	234/215

Fonte: Autoria própria

3. METODOLOGIA

O objetivo deste trabalho foi analisar a base de dados E-Saúde para identificar seu potencial de fornecer padrões úteis para gestores do sistema de Saúde de Curitiba. Para tal foi realizada uma pesquisa buscando avaliar a informação que potencialmente poderia ser extraída dos dados da base E-Saúde.

Para atingir esse objetivo, foi aplicada a Mineração de Dados (MD) nos dados oriundos do E-Saúde. Tal base de dados passou por um processo de importação, limpeza, transformação, seleção e validação de informações.

Parte do trabalho foi desenvolvida, quando possível, seguindo os passos do KDD (FAYYAD et al., 1996a). Foi necessário adaptar os passos do KDD ao longo do processo, em função das características da base de dados utilizada.

Cada fase é descrita aqui de forma pormenorizada devido a sua importância no processo como um todo. Foram descritos detalhes e decisões de cunho prático que usaram experiência técnica do autor na manipulação e construção de artefatos tecnológicos utilizados.

A limpeza dos dados (Fase 1) foi realizada durante o processo de importação dos dados para uma tabela do banco de dados. A integração de dados (Fase 2) não foi necessária, pois todos os dados vieram de uma mesma fonte. Houve então uma inversão na ordem dos passos do KDD, pois foi executada a transformação de dados (Fase 4), na qual foram criados novos atributos, calculados ou inferidos em função dos existentes. No próximo foi feita a seleção dos atributos (Fase 3) que seriam efetivamente utilizados.

Na fase de mineração dos dados (Fase 5) foram aplicados algoritmos classificadores e o processo foi encerrado com a avaliação e representação do conhecimento (Fase 6). Nas próximas seções foi descrito como cada uma das fases foi realizada.

3.1. A BASE E-SAÚDE

A PMC disponibiliza, via seu Portal de Dados Abertos (PORTAL, 2018), documentos, informações e dados governamentais de domínio público para a livre utilização. Isso garante o acesso aos dados primários que podem ser utilizados para produzir novos serviços para a sociedade.

Bases de dados de vários órgãos da PMC estão disponíveis para download em formato aberto para uso e edição irrestrita, sem a necessidade de assinatura de termo ou de identificação pessoal, com ou sem finalidade comercial. As bases de dados são atualizadas mensalmente.

No Portal de Dados Abertos estão disponíveis 20 bases de dados distribuídas entre as áreas de Recursos Humanos (1), de Finanças (2), de Abastecimento (1), de Cultura (2), do Governo Municipal (3), de Transporte (1), de Habitação (1), de Pesquisa e Planejamento (1), de Segurança (1), de Administração Pública (2), de Legislação (1) e de Saúde (4) (PORTAL, 2018).

As bases de dados (BD) da área de saúde são provenientes do sistema de informática da Secretaria Municipal de Saúde, o E-Saúde, que registra os atendimentos prestados na rede de atenção à saúde. As quatro BD disponibilizadas são: Base Médico, Base Enfermagem, Base Odontologia, Base Outros Profissionais de Nível Superior.

As BD podem ser descarregadas no formato .CSV (*Comma Separated Values*), que pode ser aberto e lido por qualquer editor de texto simples. Neste formato, cada linha do arquivo texto funciona como a representação de uma linha em uma planilha e cada célula da linha, que contém uma unidade de informação, é separada por uma vírgula.

Após uma inspeção preliminar nos dados de cada uma das quatro BD relacionadas a saúde, percebeu-se que a Base Odontologia continha apenas algumas centenas de atendimentos registrados. Estimou-se que esse número parecia não corresponder aos atendimentos realizados da rede municipal, por este motivo essa base não foi considerada neste estudo.

A Base Outros Profissionais de Nível Superior também não foi considerada neste estudo. Embora houvesse registros na ordem das dezenas de milhares, eles representavam majoritariamente o encaminhamento de médicos para estes profissionais.

Assim, ficaram as bases de Enfermagem e Médicos. Ambas com centenas de milhares de registros abrangendo toda a Rede de Atenção do município. Para esse trabalho foi selecionada a base que contém os dados referentes ao perfil de atendimento dos profissionais médicos das UBS. Esta opção foi tomada, porque é na base Médico que estão registrados os CIDs correspondentes aos atendimentos clínicos realizados.

O arquivo da Base Médico é disponibilizado todos os meses, mas contém os dados consolidados dos últimos três meses. Os dados são referentes a toda rede de atendimento no primeiro trimestre de 2017, período de tempo considerado suficiente para se obter um panorama dos atendimentos realizados. O arquivo possui tamanho de 280 MB, totalizando 730.738 linhas (instâncias).

Na mesma página do portal, está disponível o dicionário de dados da base. Este dicionário contém descrição dos atributos, bem como outros detalhes técnicos como o tipo de atributo (inteiro, decimal, string) e o seu tamanho. Na Base Médico são disponibilizados 37 atributos para cada instância.

3.2. IMPORTAÇÃO DOS DADOS

Para realizar a tarefa de importação de dados foi desenvolvida uma aplicação usando a linguagem de programação PHP. Ela lê a base de dados e em seguida grava os dados em uma tabela do banco de dados MySQL.

O PHP (um acrônimo recursivo para *PHP: Hypertext Preprocessor*) é uma linguagem de programação baseada em scripts (PHP, 2018). Ela é *open source*, de uso geral, muito utilizada, e especialmente adequada para o desenvolvimento de aplicações web, pois pode ser embutida dentro do código HTML (BENTO, 2013).

3.2.1. A Aplicação PHP

Para importação de dados foi desenvolvido, na forma de uma aplicação web, um extrator de dados de arquivos .CSV. Ele é composto de funções específicas para abertura de base de dados, leitura da instância (atributo a atributo), gravação de instâncias no banco de dados e exibição de dados estatísticos do processo de extração e dos dados do banco MySQL. Foi desenvolvida também uma interface

gráfica que reportava os eventuais problemas de importação, bem como várias informações sobre os dados importados.

Como foram importados centenas de milhares de instâncias, os tempos de importação alcançaram aproximadamente 24 horas usando um microcomputador com Windows 10 instalado, com 16 GB de memória RAM e um HD SSD de 256 GB.

3.2.2. Banco de Dados MySQL

Para a realização do trabalho, optou-se por armazenar as instâncias da Base Médico em um banco de dados MySQL. Trabalhar com os dados diretamente da base de dados original, disponibilizada no formato .CSV, não se mostrou viável. Os dados continham muitos erros, valores discrepantes, além de apresentarem caracteres especiais de controle que dificultavam a sua manipulação.

O MySQL foi escolhido por ser um eficiente Sistema de Gerenciamento de Banco de Dados (SGBD) (MySQL, 2018) muito utilizado em aplicações comerciais e de pesquisa. Seu sucesso deve-se em grande medida à fácil integração com o PHP, ambos eram incluídos nos mesmos pacotes de desenvolvimento (BENTO, 2013).

O MySQL, como as BD em geral, facilita a realização de consultas e pesquisas (TRONCHONI et al., 2010), principalmente por possuir uma linguagem para consulta e manipulação de dados, o SQL (*Structured Query Language*). Outra vantagem é a estrutura relacional do MySQL, que é mais adequada para o armazenamento dos atributos de interesse e suas características como tipo e tamanho.

Usando o MySQL, foi criada uma base de dados específica para o trabalho chamada de MESTRADO, na qual foi criada a Tabela MEDICO, com os 37 atributos da base de dados, de acordo com o descrito no dicionário de dados.

3.2.3. Limpeza dos Dados

Simultaneamente às atividades de extração e importação foi realizada a de limpeza dos dados ou pré-processamento. No início, a extração foi realizada de acordo com os atributos descritos no dicionário de dados. Isso foi feito para confirmar se os tipos dos dados recuperados da base correspondiam com os tipos

indicados pelo dicionário de dados. Nos primeiros registros recuperados, pode-se verificar que os tipos de dados extraídos estavam de acordo com o esperado.

Em seguida, obteve-se uma amostra da base de dados, com pouco mais de 1% do volume total dos registros. Isso foi feito com o intuito de se identificar o conteúdo de cada um dos atributos da base de dados.

Nesta fase exploratória foram produzidas estatísticas descritivas da amostra. Para cada registro, foram levantadas todas as diferentes ocorrências de conteúdo, bem como, a não existência de informação (se o atributo estava vazio).

Com estas estatísticas preliminares, apenas da amostra, foi possível ter noção geral de todos os atributos da base E-Saúde. Isso guiou a próxima atividade, que consistiu em extrair todo o restante dos dados.

Muitas vezes o processo de importação avançava até encontrar um problema que o obrigava parar. A cada parada eram verificados os *logs* que a aplicação PHP exibia na tela. A análise dos *logs* levou a ajustes nas funções de importação, tornando-as maduras e resistentes aos problemas da base de dados.

A cada parada, a Tabela MEDICO era apagada e recriada, fazendo com que os procedimentos de importação fossem aplicados ao conjunto completo de dados, garantindo que todos os dados seriam processados da mesma forma.

Na fase de limpeza dos dados foram corrigidas inconsistências, pois os conteúdos dos atributos podiam estar incompletos, redundantes, ruidosos e esparsos (COSTA et al., 2018). A limpeza garante a confiabilidade dos dados que serão utilizados na fase da mineração, pois caso não sejam corrigidos, contornados ou minimizados, eventuais erros nos dados podem comprometer a eficácia da mineração de dados (HAN et al., 2011).

No quadro 2 podem ser vistas as principais causas das paradas ocorridas no processo de importação e as ações tomadas.

Quadro 3 - Principais problemas encontrados durante a limpeza dos dados.

Descrição do problema	**Ação realizada**
Atributos vazios	Foi inserido o caracter “?”
Valores inteiros negativos, para atributos que deveriam ser positivos	Foram ajustados de acordo com o que se esperava do atributo
Caracteres de controle invisíveis	Foram retirados e o texto ajustado
Caractere apóstrofe em nomes	Foram retirados ajustando o nome

Ocorrência de pontos e vírgulas em textos	Substituídos por espaços

Fonte: Autoria própria.

Para cada problema, correções e ajustes foram realizados, a fim de que os atributos pudessem ser importados. Desta forma, nenhuma instância foi eliminada.

Para facilitar a posterior montagem dos arquivos (.ARFF), que seriam utilizados na fase da mineração de dados, toda vez que um atributo tinha valor nulo, foi inserido o caractere “?” no respectivo atributo.

3.2.4. Dados Importados

Na MD, os atributos podem ser divididos em dois grandes grupos: os contínuos e os nominais. Atributos representados usando números reais são contínuos. Atributos nominais podem ser uma descrição, categoria ou um nome (VASCONCELOS, 2002). Atributos nominais também podem ser chamados de atributos categóricos.

Com relação aos tipos de dados dos atributos da Tabela MEDICO, a maioria deles (30) são strings. Os atributos numéricos são quatro, sendo todos inteiros, não existem atributos com números reais. Existem ainda dois atributos do tipo data.

Com o processo de importação de dados finalizado, foi possível analisar a completude dos dados. Nessa análise foram computadas a porcentagem de ocorrências em que o atributo foi preenchido e a quantidade de valores distintos para cada atributo (Tabela 4), para cada um dos 37 atributos.

Tabela 4 - Atributos da base com informações dos atendimentos médicos.

	Descrição do Atributo	Ocorrência Percentual	Valores Distintos
1	Data de Realização do Atendimento	100,00	93
2	Data de Nascimento do Paciente	100,00	33449
3	Sexo do Paciente	100,00	2
4	Código do Tipo de Unidade de Atendimento	100,00	3
5	Tipo de Unidade de Atendimento	100,00	3
6	Código da Unidade de Atendimento	100,00	126
7	Descrição da Unidade de Atendimento	100,00	126
8	Código do Procedimento Realizado	100,00	20
9	Descrição do Procedimento Realizado	99,90	20
10	Código da Ocupação do Profissional	100,00	23
11	Descrição da Ocupação do Profissional	100,00	23

12	Código do Diagnóstico	100,00	5071
13	Descrição do Diagnóstico	99,97	5070
14	Indica se Ocorreu Solicitação de Exames	100,00	2
15	Quantidade de Medicamentos Prescritos na Farmácia Curitibana	24,45	5016
16	Quantidade de Medicamentos Dispensados na Farmácia Curitibana	12,25	984
17	Quantidade de Medicamentos Não Padronizados	13,70	1165
18	Indica se Houve Encaminhamento para Atendimento de Especialista	100,00	2
19	Área de Atuação	7,55	96
20	Indica se Desencadeou Internamento	100,00	2
21	Data do Internamento do Paciente	0,89	97
22	Estabelecimento que Solicitou o Internamento	0,89	101
23	Estabelecimento que Houve a Internação	0,89	29
24	Código do Diagnóstico do Internamento	0,89	750
25	Tipo de Tratamento de Água no Domicílio	86,97	5
26	Tipo de Abastecimento de Água no Domicílio	86,97	6
27	Indica se Há Energia Elétrica no Domicílio	100,00	2
28	Tipo de Habitação no Domicílio	86,98	9
29	Destino do Lixo no Domicílio	86,98	5
30	Destino das Fezes/Urina no Domicílio	86,98	7
31	Quantidade de Cômodos no Domicílio	100,00	97
32	Serviços Procurados em Caso de Doença	86,96	156
33	Grupo Comunitário em que o Paciente Participa	86,94	6
34	Meios de Comunicação Utilizados no Domicílio	86,96	8
35	Meios de Transporte Utilizados no Domicílio	86,96	49
36	Município do Paciente	100,00	291
37	Bairro do Paciente	100,00	1024

Fonte: Autoria própria.

3.2.5. Análise dos Dados Importados

Após o processo de importação da Tabela MEDICO, foi possível realizar análises para o entendimento das características referentes à natureza dos atributos da tabela e da base E-Saúde.

Primeira análise: os atributos da base E-Saúde podem ser divididos em quatro grupos relacionados pela similaridade do assunto tratado. São eles: Grupo Consulta do Paciente, Grupo Farmácia Curitibana, Grupo Internação do Paciente e Grupo Socioeconômico. Os atributos de cada um dos grupos são apresentados nos Quadros 4, 5, 6 e 7.

Quadro 4 - Grupo Consulta do Paciente.

Atributos do Grupo Consulta do Paciente

- Data de Realização do Atendimento
- Código do Tipo de Unidade de Atendimento
- Tipo de Unidade de Atendimento
- Código da Unidade de Atendimento
- Descrição da Unidade de Atendimento
- Código do Procedimento Realizado
- Descrição do Procedimento Realizado
- Código da Ocupação do Profissional
- Descrição da Ocupação do Profissional
- Código do Diagnóstico
- Descrição do Diagnóstico
- Indica se ocorreu solicitação de Exames
- Indica se Houve Encaminhamento para Atendimento de Especialista

Fonte: Autoria própria.

Quadro 5 - Grupo Farmácia Curitibana.

Atributos do Grupo da Farmácia Curitibana

- Quantidade de medicamentos prescritos na Farmácia Curitibana
- Quantidade de medicamentos dispensados na Farmácia Curitibana
- Quantidade de Medicamento Não Padronizado

Fonte: Autoria própria.

Quadro 6 - Grupo Internação do Paciente.

Atributos do Grupo da Internação do Paciente

- Indica se desencadeou Internamento
- Data do Internamento do paciente
- Estabelecimento que solicitou o internamento
- Estabelecimento que houve a internação
- Código do diagnóstico do internamento

Fonte: Autoria própria.

Quadro 7 - Grupo Sócio Econômico.

Atributos do Grupo Sócio Econômico

- Data de Nascimento do Paciente

- Sexo do Paciente
- Tipo de Tratamento de Água no domicílio
- Tipo de Abastecimento de Água no domicílio
- Indica se há energia elétrica no domicílio
- Tipo de habitação no domicílio
- Destino do lixo no domicílio
- Destino das fezes/urina no domicílio
- Quantidade de Cômodos no domicílio
- Serviços procurados em caso de doença
- Grupo Comunitário em que o paciente participa
- Meios de Comunicação utilizados no domicílio
- Meios de Transporte utilizados no domicílio
- Município do paciente
- Bairro do paciente

Fonte: Autoria própria.

Segunda análise: foi realizada a análise da completude da base E-Saúde. Foram computados índices de preenchimento para cada um dos atributos dos quatro grupos definidos na primeira análise. Os atributos do grupo Consulta do Paciente estão quase completos com um índice de aproximadamente 100% para todos os atributos.

Os atributos do grupo Socioeconômico apresentam um índice de preenchimento entre 86-100%. Em quatro dos treze atributos que compõe esse grupo, o preenchimento foi completo. O índice de preenchimento pode ser considerado alto, pois nem todos esses atributos são de preenchimento obrigatório.

Já os atributos do grupo da Farmácia Curitibana apresentaram um índice de preenchimento baixo, entre 12% e 24%. Além disso, foi identificada uma quantidade significativa de dados incoerentes. Por exemplo, em uma mesma consulta teriam sido dispensados 1000 medicamentos a um paciente. Em função disso, foram importadas apenas as instâncias com informações dentro de um limite considerado razoável.

Por fim, o grupo Internação do Paciente apresentou o mais baixo índice de preenchimento de 0,82%. Mas isso era esperado, pois somente quando ocorrem internações é que os atributos desse grupo são preenchidos.

O índice de preenchimento dos atributos é relevante, pois a mineração de dados pode ser comprometida se o índice for baixo. De forma geral, na base E-Saúde, a grande maioria dos atributos é completa o suficiente para as tarefas de classificação. A maioria dos atributos pouco preenchidos são aqueles cujo preenchimento não é obrigatório ou o preenchimento é condicional.

Terceira análise: foi verificada a existência de pares de atributos que descrevem aspectos diferentes da mesma informação. Esses pares são compostos por um atributo com um código e o outro com a descrição do item representado pelo código. Do ponto de vista da mineração de dados apenas um deles é necessário. No Quadro 8 são mostrados os pares identificados na Tabela MEDICO.

Quadro 8 - Atributos pares de código e descrição.

Atributo Código	Atributo Descrição
Código do Tipo de Unidade de Atendimento	Tipo de Unidade de Atendimento
Código da Unidade de Atendimento	Descrição da Unidade de Atendimento
Código do Procedimento Realizado	Descrição do Procedimento Realizado
Código da Ocupação do Profissional	Descrição da Ocupação do Profissional
Código do Diagnóstico	Descrição do Diagnóstico

Fonte: Autoria própria.

3.3. CRIAÇÃO DE NOVOS ATRIBUTOS

No processo de mineração de dados, a criação de atributos a partir de outros existentes é feita de modo que informações importantes sejam extraídas de um conjunto de atributos de forma mais eficaz (COSTA, 2007). Para este trabalho, dez novos atributos foram criados. O objetivo foi melhorar a efetividade dos atributos da base E-Saúde.

Os atributos criados foram acrescentados à Tabela MEDICO por meio de scripts em SQL. Os valores destes atributos foram computados a partir dos atributos originais, utilizando-se comandos SQL para isso.

No Quadro 9 estão descritos os novos atributos criados, os atributos originais que foram utilizados na criação deles e o grupo ao qual cada um foi incluído.

Quadro 9 - Atributos criados.

	Atributo Novo	Atributo Original	Grupo
1	Dia da Semana do Atendimento	Data de Realização do Atendimento	Consulta Paciente
2	Idade do Paciente	Data de Nascimento do Paciente	Consulta Paciente
3	Faixa de Idade do Paciente	Idade do Paciente	Consulta Paciente
4	Capítulo do CID do Diagnóstico	Código do Diagnóstico	Consulta Paciente
5	Descrição do Capítulo do CID	Descrição do Diagnóstico	Consulta Paciente
6	Prescrito da Farmácia Curitibana	Quantidade de medicamentos prescritos na Farmácia Curitibana	Farmácia Curitibana

7	Dispensado da Farmácia Curitibana	Quantidade de medicamentos dispensados na Farmácia Curitibana	Farmácia Curitibana
8	Medicamentos Não Padronizados	Quantidade de Medicamento Não Padronizado	Farmácia Curitibana
9	Faixa de Cômodos da Casa do Paciente	Quantidade de Cômodos no domicílio	Sócio Econômico
10	Origem em Curitiba	Município do paciente	Sócio Econômico
11	Resultado do atendimento	Indica se Houve Encaminhamento para Atendimento de Especialista Indica se ocorreu solicitação de Exames Indica se desencadeou Internamento	Consulta Paciente

Fonte: Autoria própria.

3.4. SELEÇÃO DE ATRIBUTOS

Na fase de importação, todos os atributos foram importados na tabela SQL e nenhum tipo de pré-seleção foi realizada. Contudo, é necessário para o sucesso da mineração de dados, que sejam aplicados métodos de tratamento, limpeza e redução do volume de dados.

A seleção de atributos é uma técnica utilizada com o intuito de reduzir a quantidade dos dados, facilitando a aplicação de algoritmos de mineração. Esta redução visa eliminar atributos que não agregam informações para a análise, produzindo assim uma representação mais compacta, mais facilmente interpretável do objetivo a ser alcançado, focalizando a atenção do usuário sobre os atributos mais relevantes (WITTEN et al., 2011).

Atributos que não contribuem são os que não pertencem ao escopo da análise ou que possuem valores repetidos com alta frequência. Atributos com uma alta concentração de um valor tendem a ser irrelevantes ou mesmo prejudiciais aos algoritmos de mineração (FERREIRA, 2015).

A natureza dos atributos da base E-Saúde mostrou que era necessário reduzir sua quantidade. A seleção dos atributos foi realizada de forma manual e em etapas, utilizando o critério da relevância que o atributo poderia ter na mineração dos dados.

Inicialmente foram retirados os atributos de código e mantidos os com descrições legíveis nas análises. Devido à baixa frequência, menor que 1% dos registros, foram descartados os atributos do grupo Internação do Paciente, com exceção do atributo *Indica se desencadeou Internamento* que estava presente em 100% dos registros.

Do conjunto formado pelos 37 atributos originais importados e pelos 10 atributos criados para melhor compreensão dos dados, foram considerados relevantes para a análise 26 atributos, descritos na Tabela 5. Os atributos descritos em caixa baixa são os atributos originais da base, já os em caixa alta são atributos criados. Também são exibidos os nomes e as descrições dos atributos da Tabela MEDICO e dos arquivos .ARFF usados nos experimentos.

Tabela 5 - Atributos Selecionados para Mineração

	Atributo	Descrição do Atributos
1	DIA_DA_SEMANA	Dia da Semana do Atendimento
2	IDADE_FAIXA	Faixa da Idade do Paciente
3	sexo	Sexo do Paciente
4	descr_unidade	Descrição da Unidade de Atendimento
5	descr_procedimento	Descrição do Procedimento Realizado
6	descr_CBO	Descrição da Ocupação do Profissional
7	cod_CID	Código do Diagnóstico
8	CAP_CID	Capítulo do CID do Diagnóstico
9	solicit_exames	Indica se Ocorreu Solicitação de Exames
10	FARM_PRESCR	Prescrito da Farmácia Curitibana
11	FARM_DISPEN	Dispensado da Farmácia Curitibana
12	FARM_NAO_PADRON	Medicamentos Não Padronizados
13	enc_atend_especia	Indica se Houve Encaminhamento para Atendimento de Especialista
14	desencadeou_interna	Indica se Desencadeou Internamento
15	tratam_domic	Tipo de Tratamento de Água no Domicílio
16	abastecimento_agua	Tipo de Abastecimento de Água no Domicílio
17	energia_eletrica	Indica se Há Energia Elétrica no Domicílio
18	tipo_habitacao	Tipo de Habitação no Domicílio
19	Destino Lixo	Destino do Lixo no Domicílio
20	fezes_urina	Destino das Fezes/Urina no Domicílio
21	COMODOS_FAIXA	Faixa de Cômodos da Casa do Paciente
22	em_caso_doenca	Serviços Procurados em Caso de Doença
23	grupo_comunitario	Grupo Comunitário em que o Paciente participa
24	meio_comunicacao	Meios de Comunicação Utilizados no Domicílio
25	meio_transporte	Meios de Transporte Utilizados no Domicílio
26	ORIGEM_CURITIBA	Origem Em Curitiba

Fonte: Autoria própria.

3.5. O WEKA

A recuperação convencional de informações por consultas SQL não permite explorar o potencial de uma grande massa de dados como a base de dados E-

Saúde (COSTA et al., 2018). A mineração de dados, por sua vez, permite investigar o conjunto dos dados em busca de padrões que tenham valia.

Pode-se utilizar diferentes tipos de algoritmos de MD para extrair padrões dos dados, com cada tipo fornecendo respostas diferentes para um mesmo problema de entrada. Por isso, um prévio conhecimento do problema e do que se espera como resposta, influencia diretamente na escolha da tarefa a ser executada, bem como o algoritmo, ou algoritmos a serem utilizados.

Neste trabalho, a extração de padrões foi realizada por algoritmos que fazem classificação, tendo em vista que a saída de um classificador formata o conhecimento de forma textual, o que pode facilitar sua compreensão para os gestores na área de saúde.

Existe uma grande quantidade de ferramentas de mineração de dados, tanto comerciais quanto *open source*. Dentre elas, o Weka (*Waikato Environment for Knowledge Analysis*) foi selecionado por ser reconhecido como referência em mineração de dados (WEKA, 2018).

O Weka foi desenvolvido pela Universidade de Waikato, na Nova Zelândia, em linguagem JAVA, na forma de uma coleção de algoritmos de aprendizado de máquina para a realização de tarefas de Mineração de Dados.

Ele tem como principal objetivo resolver problemas de mineração de dados baseado em aprendizagem de máquina, por meio de métodos de análise em conjuntos de dados, de uma forma rápida e flexível (WITTEN et al., 2011). Os motivos que levaram a escolha do Weka são:

- É um software livre, que não exige o pagamento de licenças. Também pode ser descarregado e instalado com facilidade.
- Possui uma interface gráfica que torna o aprendizado e posterior uso mais fácil e intuitivo.
- É uma ferramenta madura, desenvolvida ainda no começo dos anos 90, que possui um conjunto abrangente de algoritmos que executam diferentes tarefas de aprendizado de máquina.

O Weka Explorer é uma interface gráfica de uso intuitivo, implementada de modo a usar as bibliotecas do Weka (WEKA EXPLORER, 2018). A Figura 8 mostra a tela do Weka Explorer, carregado com os dados do Distrito Sanitário do Boqueirão.

Cada um dos principais pacotes do Weka - Filtros, Classificadores, Clusterers, Associações e Seleção de Atributo - é disponibilizado no Explorer juntamente com uma ferramenta de visualização, que permite que os conjuntos de dados e as previsões de Classificadores e Clusterers sejam visualizados em duas dimensões.

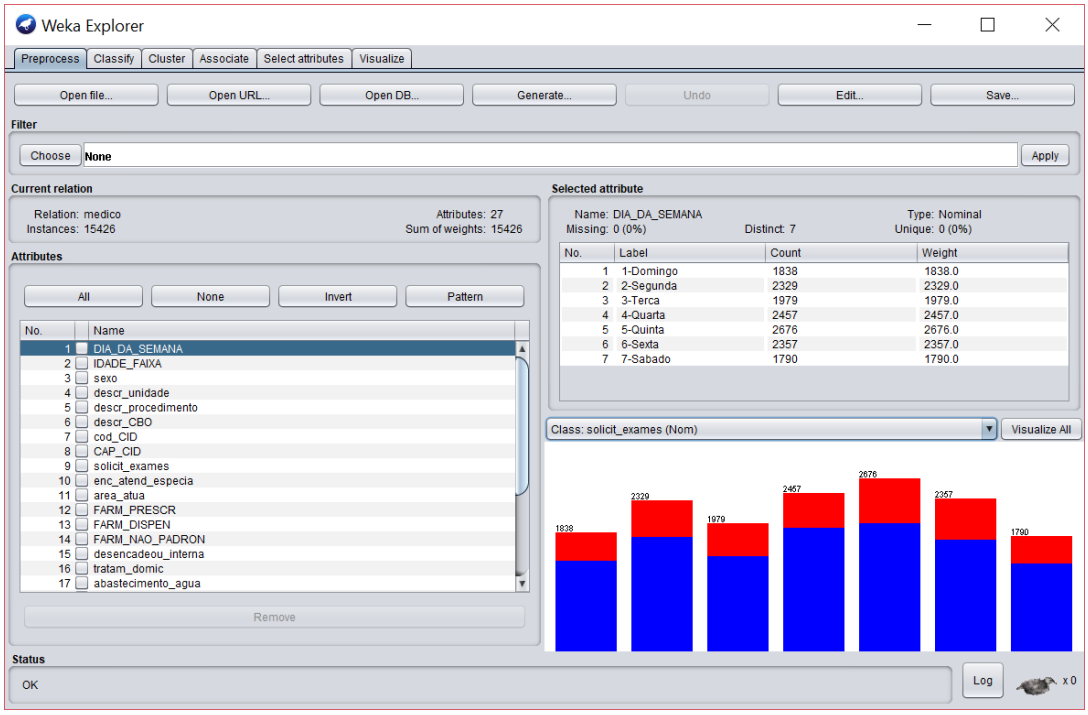


Figura 8 - Weka Explorer com dados Distrito Sanitário do Boqueirão.

Fonte: Autoria Própria

3.5.1. O Arquivo .ARFF

O arquivo do tipo .ARFF (*Attribute-Relation File Format*) foi desenvolvido pelo *Machine Learning Project* no Departamento de Ciência da Computação da Universidade de Waikato para ser usado com o software de aprendizado de máquina Weka (ARFF, 2018). O .ARFF é o formato padrão de armazenamento de dados no Weka. O Weka também aceita outros tipos de entrada de dados, como arquivos .CSV ou conexões diretas com banco de dados.

O arquivo .ARFF é um arquivo de texto ASCII que descreve uma lista de instâncias que compartilham um conjunto de atributos. Ele possui duas seções distintas. A primeira seção contém as informações do cabeçalho, no qual estão os metadados referentes aos dados armazenados no arquivo. A segunda seção contém todos os dados de cada instância (Quadro 10).

```
% Creator: Santos, William

@relation medico

@attribute sexo {F,M}
@attribute cod_CBO {225124,225125,225142,225250}
@attribute solicit_exames {Nao,Sim}
@attribute enc_atend_especia {Nao,Sim}
@attribute desencadeou_interna {Nao,Sim}
@attribute ORIGEM_CURITIBA {N,S}

@data
M,225125,Nao,Nao,Nao,N
F,225125,Nao,Nao,Nao,?
F,225250,Nao,Nao,Nao,S
F,225142,Nao,Sim,Nao,S
M,225124,Nao,Nao,Nao,S
F,225142,Sim,?,Nao,S
F,225142,Nao,Nao,Nao,S
```

Quadro 10 – Arquivo .ARFF com atributos da base E-Saúde

Fonte: Autoria própria.

O cabeçalho contém o nome da relação (`@relation medico`), uma lista dos atributos com as respectivas faixas de valores que podem ser a ele atribuídos. A seção com os dados começa com a diretiva `@data` e constitui a maior parte do arquivo .ARFF.

É importante destacar que o ponto de interrogação “?” indica um valor desconhecido ou faltante. Assim, um ponto de interrogação foi inserido na Tabela MEDICO, no respectivo atributo, durante o processo de importação, cada vez que uma informação no arquivo .CSV original não estava presente.

A criação dos arquivos .ARFF para o Weka foi feita pela mesma aplicação PHP que realizou a importação dos dados. A aplicação PHP recebeu duas novas páginas com funções para seleção de atributos da Tabela MEDICO e a criação propriamente dita dos arquivos .ARFF.

3.5.2. Algoritmos Classificadores J48 e JRip

Para realização do trabalho foram escolhidos os algoritmos classificadores J48 e JRip, um indutor de Árvore de Decisão e outro baseado em Regras, disponíveis no software Weka. Ambos algoritmos, com paradigma simbólico (SOUZA e FELIPPO, 2018). O principal motivo dessa escolha é o fato de que as saídas

desses algoritmos serem na forma textual, o que permite que as regras sejam facilmente lidas por pessoas não especialistas, como, por exemplo, gestores de saúde. Detalhes técnicos de cada classificador podem ser encontrados na seção de Anexos.

O algoritmo J48 é uma implementação em Java do classificador C4.5 que pode induzir uma Árvore de Decisão podada ou não. Ele é um algoritmo versátil, que classifica tanto atributos nominais quanto numéricos e tem como base o algoritmo de Hunt, que também é a base muitos outros algoritmos de indução de Árvores de Decisão, incluindo o ID3, CART e o C4.5 (SOUZA, 2011).

O algoritmo J48 trata bases com dados ausentes evitando assim a necessidade de se discretizar atributos numéricos e inserir valores ausentes (QUINLAN, 1993). Esse algoritmo é adequado para uma base complexa como a do E-Saúde que contém atributos nominais (em sua maioria), numéricos e ausentes.

O JRip foi utilizado como indutor de um classificador baseado em regras. Ele é uma versão otimizada do algoritmo IREP, implementada na linguagem Java. O algoritmo classificador JRip implementa o *Repeated Incremental Pruning to Produce Error Reduction*, ou RIPPER, que significa Poda Incremental Repetida para Produzir Redução de Erro (COHEN, 1995).

O JRip é um algoritmo baseado em regras de decisão, que analisa o conjunto de instâncias e gera regras no formato lógico, as quais comumente combinam atributos para capturar mais adequadamente as classes de instâncias. Seu funcionamento está baseado primeiro na escolha da classe majoritária como padrão de comparação. Em uma segunda fase, o algoritmo tenta descobrir as regras para detectar as classes minoritárias, otimizando o conjunto de regras iniciais para diminuir erros e tornar o processo mais seletivo.

Este algoritmo é especialmente apropriado para a indução de modelos a partir de conjuntos de dados com distribuição de classes desequilibradas. Também funciona bem com conjuntos de dados com ruídos porque usa um conjunto de validação para evitar o *overfitting* do modelo (REZENDE, 2005).

4. EXPERIMENTOS, RESULTADOS E DISCUSSÃO

O total de instâncias nas bases de dados com os atendimentos médicos, de enfermagem e de profissionais de nível superior foi de 1.468.466, ou um total de 578 MB de informação.

Inicialmente, foi realizada uma série de experimentos, com o intuito de se identificar os atributos mais promissores a serem classificados dentre os atributos constituintes da base de Dados E-Saúde.

Em função dos resultados encontrados, ou seja, do conhecimento apresentado quando se utilizava um atributo como meta (objetivo) de um classificador, destacaram-se seis atributos como possíveis de serem estudados em maior profundidade.

São eles: `Cod_CID`, `CAP_CID`, `solicit_exames`, `enc_atend_especia`, `desencadeou_interna` e `ORIGEM_CURITIBA`. A partir desta seção, os atributos serão referenciados por seus nomes, não mais pela sua descrição.

Analisando os dados, percebeu-se que eles não poderiam ser tratados juntos, pois estavam conceitualmente misturados. Os arquivos disponibilizados no Portal de Dados Abertos consolidam as instâncias em um único conjunto, mas para a melhor aplicação da MD dessas instâncias é necessário dividi-las em subgrupos.

A aplicação da MD no conjunto completo não resultou na descoberta de conhecimento relevante. Por exemplo, um resultado preliminar consolidado, foi que o fato do paciente morar em Curitiba implicaria em ele não ser internado.

Analisando novamente os atributos, percebeu-se que os atendimentos aos pacientes deveriam ser separados pelo tipo de profissional que proveu o atendimento: médicas (os), enfermeiras (os) ou outras (os) profissionais de nível superior.

Percebeu-se, também, que o tipo de unidade onde foram realizados os atendimentos era relevante. Assim, os atendimentos deveriam ser divididos pelo tipo de unidade em que foram realizados: nas Unidades Básicas de Saúde, nas Unidades de Pronto Atendimento ou em um Centro de Especialidades.

Considerando esses dois aspectos, o tipo de profissional que realizou o atendimento e o tipo de unidade onde ele foi realizado, os dados foram

segmentados. Esta segmentação é apresentada no Quadro 10, no qual estão mostrados os nove segmentos possíveis de atendimento a um paciente que acessa a Rede de Atenção de Curitiba.

Quadro 10 - Local x Tipo de Atendimento.

Atendimentos	**Tipo de Unidade**		
	UBS	**UPA**	**Especialidades**
Médico	UBSxMed	UPAxMed	EspxMed
Enfermeiro	UBSxEnf	UPAxEnf	EspxEnf
Profissional de Nível Superior	UBSxSup	UPAxSup	EspxSup

Fonte: Autoria própria.

Os resultados dos experimentos preliminares para explorar as potencialidades da base de dados E-Saúde foram apresentados a quatro servidores da Rede Municipal de Saúde de Curitiba que trabalhavam em três Unidades Básicas de Saúde e uma Unidade de Pronto atendimento.

Um servidor e uma servidora desempenhavam funções de nível técnico e duas servidoras de nível superior. Sendo que essas duas são especialistas em suas áreas de formação e as responsáveis administrativas pelas suas unidades. Elas fizeram sugestões que contribuíram no direcionamento da próxima etapa do trabalho, mais especificamente elas indicaram casos sobre os quais, elas julgaram, seria interessante obter mais informações.

Para descobrir o potencial da base de dados em fornecer conhecimento, decidiu-se então analisar a resolutividade dos atendimentos realizados nas unidades de saúde, mais especificamente aquela relativa aos pedidos de solicitações de exames pelos médicos nas consultas realizadas.

Optou-se por estudar os pedidos de solicitações de exames, pois eles ocorrem frequentemente. Informação sobre o que está acontecendo na geração destas solicitações pode fornecer subsídios para que os gestores tomem medidas que potencialmente levem à economia ou à otimização dos recursos. Também este estudo foi realizado na base dos MÉDICOS, porque são os médicos que efetivamente fazem as solicitações dos exames.

Além disso, as gestoras consultadas também sugeriram delimitar a análise às unidades de saúde constituintes de um mesmo distrito sanitário da cidade, onde seria encontrada uma população com características epidemiológicas e sociais semelhantes (CARDOSO et al., 2013).

Foram escolhidos dois distritos, o do Boqueirão e o do Cajuru, por terem um número de equipamentos e população atendida relativamente semelhantes. Também foram escolhidos dois distritos para que se pudesse observar e comparar os resultados obtidos.

Para a realização dos experimentos, foram selecionados os atendimentos médicos das Unidades Básicas de Saúde e a Unidade de Pronto Atendimento de ambos os distritos.

Nos experimentos, em todos os treinamentos e testes realizados, foi utilizada a opção de *cross-validation* utilizando 10 *folds*, no intuído de estimar de forma mais realista as taxas de acertos/erros de classificação. Foram aplicados os algoritmos classificadores J48 e JRip tendo como atributo a ser classificado o Solicitação de Exames.

O WEKA foi alimentado por dois arquivos, um com os dados do DS do Boqueirão, e outro com os dados do DS do Cajuru, com 44.486 e 44.140 instâncias respectivamente. Ambos contendo os atendimentos médicos dos três primeiros meses de 2017.

Embora significativa, a fase de mineração de dados não é a etapa final KDD. É preciso proceder com a identificação e síntese das relações descobertas, que são comumente chamadas de padrões (SOCZEK e ORLOVSKI, 2014). Foi no momento da avaliação dos padrões identificados que eles foram interpretados. Foram, então, selecionados os padrões relevantes para o problema pesquisado, ou seja, aqueles que poderiam auxiliar as gestoras em seu planejamento ou tomada de decisão.

4.1. ANÁLISE DA RESOLUTIVIDADE DOS ATENDIMENTOS

De acordo com o Ministério da Saúde a resolutividade pode ser entendida como a exigência de que, quando um indivíduo procura um atendimento, o serviço correspondente esteja capacitado para atendê-lo e resolvê-lo até o nível da sua competência (DEGANI, 2002). A resolutividade dos serviços de saúde é uma maneira de avaliar os serviços a partir dos resultados obtidos pelo atendimento ao usuário.

Na base de dados E-Saúde, os atributos que descrevem se um paciente teve seu caso resolvido na consulta (em ordem da maior para a menor frequência de

ocorrência) são: Indica se Ocorreu Solicitação de Exames (`solicit_exames`), Indica se Houve Encaminhamento para Atendimento de Especialista (`enc_atend_especia`) e Indica se Desencadeou Internamento (`desencadeou_interna`). Se os três atributos contiverem o valor `Não`, do ponto de vista daquela consulta, o atendimento teve resolução.

Na tabela 6 são mostrados os percentuais de ocorrência na base de dados médico de cada um dos atributos que contribuem para a resolução (ou não) da consulta do paciente na unidade de saúde.

Observando os percentuais, percebe-se que 70% dos atendimentos realizados, do ponto de vista do atendimento, tem resolução, ou seja, o paciente não precisa acessar a unidade de saúde novamente.

Dos três motivos que levam o paciente a retornar a unidade de saúde ou a um outro equipamento da rede de atenção (solicitação de exames, consulta a especialistas ou solicitação de internação), são as solicitações dos exames, que mais ocorrem, correspondendo a 15% do total de atendimentos.

Tabela 6 - Estatísticas da Resolutividade dos Atendimentos Realizados

Linha	desencadeou_interna	enc_atend_especia	solicit_exames	Ocorrência %
1	N	N	N	70,136
2	N	N	S	15,977
3	N	S	N	4,902
4	N	S	S	8,099
5	S	N	N	0,374
6	S	N	S	0,511
7	S	S	N	0,001
8	S	S	S	0,001

Fonte: Autoria própria.

Em seguida, foi utilizada a mineração de dados usando o atributo `solicit_exames`. Para isso, foram aplicados os algoritmos classificadores J48 e JRip.

O conjunto dos atributos utilizados na análise da Resolutividade, com suas respectivas frequências (totais de ocorrência de cada classe), pode ser visto no Apêndice.

4.1.1. Experimentos com o Classificador J48

O algoritmo de Árvore de Decisão usa uma abordagem direcionada, ou seja, o algoritmo precisa de um atributo definido pelo usuário que terá então sua classificação prevista.

Assim, os resultados obtidos do atributo `solicit_exames`, submetidos ao algoritmo classificador J48 do WEKA para o atributo para os DS do Boqueirão e Cajuru, são mostrados nos quadros 11 e 12.

Quadro 11 – DS Boqueirão - Árvore de Decisão do atributo `solicit_exames`

```
enc_atend_especia = Nao: Nao (41858/9849)
enc_atend_especia = Sim
|   descr_CBO = MEDICO_CLINICO: Sim (4411/1962)
|   descr_CBO = MEDICO_DA_ESTRATEGIA_DE_SAUDE_DA_FAMILIA
|   |   IDADE_FAIXA = 1:Bebe: Nao (156/41)
|   |   IDADE_FAIXA = 2:Crianca: Nao (301/119)
|   |   IDADE_FAIXA = 3:Adolescente: Sim (210/87)
|   |   IDADE_FAIXA = 4:AdultoJovem: Sim (1295/437)
|   |   IDADE_FAIXA = 5:Adulto: Sim (1608/482)
|   |   IDADE_FAIXA = 6:Idoso: Sim (1536/590)
|   descr_CBO = MEDICO_GINECOLOGISTA_E_OBSTETRA: Sim (1972/236)
|   descr_CBO = MEDICO_PEDIATRA: Nao (640/250)


   5.  Number of Leaves  :   10

Size of the tree :    13
```

Fonte: Autoria própria.

Quadro 12 – DS Cajuru - Árvore de Decisão do atributo `solicit_exames`

```
enc_atend_especia = Nao: Nao (43938/10133)
enc_atend_especia = Sim
|   sexo = F
|   |   descr_CBO = MEDICO_CLINICO: Sim (1620/615)
|   |   descr_CBO = MEDICO_DA_ESTRATEGIA_DE_SAUDE_DA_FAMILIA: Sim (5203/1624)
|   |   descr_CBO = MEDICO_GENERALISTA: Sim (142/47)
|   |   descr_CBO = MEDICO_GINECOLOGISTA_E_OBSTETRA: Sim (969/153)
|   |   descr_CBO = MEDICO_PEDIATRA: Nao (223/84)
|   sexo = M
|   |   IDADE_FAIXA = 1:Bebe: Nao (165/49)
|   |   IDADE_FAIXA = 2:Crianca: Nao (386/138)
|   |   IDADE_FAIXA = 3:Adolescente: Nao (225/98)
|   |   IDADE_FAIXA = 4:AdultoJovem: Sim (679/299)
|   |   IDADE_FAIXA = 5:Adulto: Sim (1017/369)
|   |   IDADE_FAIXA = 6:Idoso: Sim (1123/517)

Number of Leaves  :   12

Size of the tree :    16
```

Fonte: Autoria própria.

Embora a interpretação de uma Árvore de Decisão seja relativamente simples, a árvore pode torna-se incompreensível se for muito grande. Desta forma foram descartados os ramos menos relevantes, com menor poder preditivo. Esta ação tende a reduzir o ruído na análise.

Esse descarte foi feito usando a poda, que no algoritmo J48, consiste em ajustar o parâmetro Número Mínimo de Objetos. O resultado é a geração de um modelo mais amigável e compreensível para a visualização na forma textual. Contudo, a representação gráfica da árvore ficou prejudicada, pois os atributos e as informações ficaram sobrepostos, impedindo uma interpretação mais clara.

Neste experimento, as Árvores de Decisão obtidas têm tamanho de 10 e número de folhas igual a 13 para o DS do Boqueirão e tamanho de 12 e número de folhas igual a 16 para o DS do Cajuru (Quadros 11 e 12).

De 26 atributos, o algoritmo identificou como raiz, ou seja, como preponderante no momento de solicitar exames, o atributo `enc_atend_especia`. No segundo nível da árvore, o algoritmo identificou o atributo `descr_CBO` para o DS do Boqueirão e o atributo `sexo` no DS do Cajuru o atributo.

A acurácia apresentada pelo algoritmo classificador J48 foi de 73,94%, e de 74,60%, para os DS do Boqueirão e do Cajuru. Percentuais bem próximos e aceitáveis para considerar como válidos os modelos encontrados.

Cabe ainda destacar uma característica das árvores obtidas é a apresentação de dois números exibidos ao lado de cada folha. O primeiro número mostra o total de classificações realizadas e o segundo, o de classificações incorretas dentro do total realizado, o que representa a eficácia da classificação realizada.

5.1.1. Experimentos com o Classificador JRip

Da mesma forma que os experimentos realizados classificador J48, os dados dos dois distritos Sanitários, foram submetidos ao classificador JRip do WEKA tendo o atributo `solicit_exames`, como o atributo a ser classificado.

De 26 atributos, o classificador JRip identificou como preponderante, em ambos os distritos sanitários, para solicitação de exames o atributo `enc_atend_especia`, com presença em 20 regras das 23 encontradas para o DS do Boqueirão e presença em 12 regras das 19 encontradas para o DS do Cajuru.

Nas Tabelas 7 e 8, as regras obtidas foram colocadas em ordem decrescente de ocorrência. Estas duas tabelas possuem duas colunas: (i) a primeira contém a regra encontrada, (ii) a segunda exibe dois números: total de classificações realizadas/total de classificações incorretas.

Tabela 7 - Regras geradas com o classificador JRip para o DS do Boqueirão

Regras com solicit_exames=Sim	Ocorrências Total/Erros
(enc_atend_especia = Sim) and (sexo = F) and (IDADE_FAIXA = 6:Idoso)	1652/735
(enc_atend_especia = Sim) and (IDADE_FAIXA = 5:Adulto) and (descr_CBO = MEDICO_DA_ESTRATEGIA_DE_SAUDE_DA_FAMILIA)	1292/431
(enc_atend_especia = Sim) and (descr_CBO = MEDICO_GINECOLOGISTA_E_OBSTETRA) and (CAP_CID = 21:CapXXI)	1217/115
(enc_atend_especia = Sim) and (CAP_CID = 13:CapXIII)	963/177
6. (enc_atend_especia = Sim) and (sexo = F) and (CAP_CID = 18:CapXVIII)	757/173
(enc_atend_especia = Sim) and (sexo = F) and (descr_CBO = MEDICO_GINECOLOGISTA_E_OBSTETRA)	745/111
(enc_atend_especia = Sim) and (IDADE_FAIXA = 4:AdultoJovem) and (descr_CBO = MEDICO_DA_ESTRATEGIA_DE_SAUDE_DA_FAMILIA) and (sexo = F)	444/185
(enc_atend_especia = Sim) and (sexo = F) and (DIA_DA_SEMANA = 2-Segunda)	413/184
(enc_atend_especia = Sim) and (FARM_NAO_PADRON = NAO_PADRON_SIM) and (FARM_PRESCR = PRESCR_NAO)	332/128
(enc_atend_especia = Sim) and (IDADE_FAIXA = 5:Adulto) and (descr_unidade = UMS_MENONITAS)	288/94
(enc_atend_especia = Sim) and (descr_unidade = UMS_MENONITAS) and (descr_CBO = MEDICO_CLINICO)	270/114
(enc_atend_especia = Sim) and (CAP_CID = 18:CapXVIII)	270/95
(enc_atend_especia = Sim) and (IDADE_FAIXA = 5:Adulto) and (descr_unidade = UMS_TAPAJOS) and (sexo = F) and (FARM_PRESCR = PRESCR_NAO) and (meio_comunicacao = TELEVISAO_OU_RADIO) => solicit_exames=Sim (40.0/8.0) (enc_atend_especia = Sim) and (CAP_CID = 9:CapIX)	191/77
(descr_CBO = MEDICO_CLINICO) and (CAP_CID = 4:CapIV) and (COMODOS_FAIXA = 4pecas)	157/68
(enc_atend_especia = Sim) and (IDADE_FAIXA = 4:AdultoJovem) and (descr_CBO = MEDICO_DA_ESTRATEGIA_DE_SAUDE_DA_FAMILIA) and (descr_unidade = UMS_WALDEMAR_MONASTIER_PSF)	153/37
(enc_atend_especia = Sim) and (descr_CBO = MEDICO_CLINICO) and (IDADE_FAIXA = 5:Adulto) and (descr_unidade = UMS_SAO_PEDRO)	117/51
(descr_CBO = MEDICO_CLINICO) and (CAP_CID = 4:CapIV) and (COMODOS_FAIXA = 5pecas) and (DIA_DA_SEMANA = 5-Quinta)	55/21
(enc_atend_especia = Sim) and (descr_unidade = UMS_JARDIM_PARANAENSE_PSF) and (COMODOS_FAIXA = 0-3pecas)	53/15
(enc_atend_especia = Sim) and (em_caso_doenca = UNIDADE_DE_SAUDE) and (CAP_CID = 11:CapXI)	37/11
(enc_atend_especia = Sim) and (descr_CBO = MEDICO_CLINICO) and (CAP_CID = 14:CapXIV)	34/10
(enc_atend_especia = Sim) and (IDADE_FAIXA = 4:AdultoJovem) and (CAP_CID = 21:CapXXI) and (descr_procedimento = CONSULTA_PRE-NATAL) and (meio_comunicacao = TELEVISAO_OU_RADIO) and (DIA_DA_SEMANA = 6-Sexta)	21/1
solicit_exames=Nao	44486/10837

Fonte: Autoria própria.

Tabela 8 - Regras geradas com o classificador JRip para o DS do Cajuru

Regras com solicit_exames=Sim	Ocorrências Total/Erros
(enc_atend_especia = Sim) and (IDADE_FAIXA = 5:Adulto)	3158/1030
(enc_atend_especia = Sim) and (IDADE_FAIXA = 6:Idoso) and (sexo = F)	1686/670
(enc_atend_especia = Sim) and (sexo = F) and (IDADE_FAIXA = 4:AdultoJovem)	1559/545
(enc_atend_especia = Sim) and (sexo = F) and (descr_CBO = MEDICO_GINECOLOGISTA_E_OBSTETRA)	969/153

(enc_atend_especia = Sim) and (IDADE_FAIXA = 4:AdultoJovem)	679/299
7. (enc_atend_especia = Sim) and (sexo = F) and (CAP_CID = 13:CapXIII)	678/91
8. (enc_atend_especia = Sim) and (sexo = F) and (FARM_DISPEN = DISPEN_SIM)	566/157
(descr_unidade = UMS_TRINDADE_PSF) and (CAP_CID = 4:CapIV) => solicit_exames=Sim (156.0/52.0) (FARM_DISPEN = DISPEN_SIM) and (IDADE_FAIXA = 4:AdultoJovem) and (CAP_CID = 21:CapXXI)	539/269
(enc_atend_especia = Sim) and (sexo = F) and (descr_CBO = MEDICO_DA_ESTRATEGIA_DE_SAUDE_DA_FAMILIA) and (grupo_comunitario = NAO_PARTICIPA)	267/124
(enc_atend_especia = Sim) and (IDADE_FAIXA = 6:Idoso) and (CAP_CID = 9:CapIX) => solicit_exames=Sim (130.0/44.0) (FARM_DISPEN = DISPEN_SIM) and (IDADE_FAIXA = 4:AdultoJovem) and (descr_CBO = MEDICO_GINECOLOGISTA_E_OBSTETRA)	232/84
9. (CAP_CID = 13:CapXIII) and (FARM_DISPEN = DISPEN_SIM)	202/61
(enc_atend_especia = Sim) and (IDADE_FAIXA = 6:Idoso) and (DIA_DA_SEMANA = 5-Quinta)	182/86
(enc_atend_especia = Sim) and (CAP_CID = 18:CapXVIII)	182/69
(CAP_CID = 9:CapIX) and (descr_unidade = UMS_TRINDADE_PSF)	169/66
(enc_atend_especia = Sim) and (IDADE_FAIXA = 6:Idoso) and (CAP_CID = 13:CapXIII)	100/19
(IDADE_FAIXA = 5:Adulto) and (FARM_DISPEN = DISPEN_SIM) and (descr_unidade = UMS_ALVORADA_PSF)	96/37
solicit_exames=Nao	44140/9937

Fonte: Autoria própria.

A acurácia apresentada pelo classificador JRip foi de 74,27%, e de 74,84%, para os DS do Boqueirão e do Cajuru. Percentuais bem próximos e aceitáveis para considerar como válidos os modelos encontrados.

9.1.1. Comparação Entre Resultados dos Classificadores J48 e JRip

As acurácias obtidas pelos dois classificadores, o J48 e JRip, quando aplicados aos dados dos dois DS Sanitários, do Boqueirão e do Cajuru, foram muito próximas, conforme mostrado na tabela 9. O desempenho dos dois algoritmos de classificação foi muito similar.

Tabela 9 - Acurácia dos Classificadores J48 e JRip

Classificador	Acurácia
J48 – DS Boqueirão	73,94
J48 – DS Cajuru	74,60
JRip – DS Boqueirão	74,27
JRip – DS Cajuru	74,84

Fonte: Autoria própria.

Ambos os algoritmos classificadores indicaram o atributo `enc_atend_especia`, *"Indica se Houve Encaminhamento para Atendimento de*

Especialista", como o principal a ser considerado quando se trata de uma solicitação de exames de um paciente.

Por sua vez, o atributo `solicit_exames`, *"Indica se Ocorreu Solicitação de Exames"*, devido a grande ocorrência (em torno de 15%) é o primeiro a ser considerado quando se quer saber se o paciente teve ou não seu caso resolvido no acesso a uma Unidade Básica de Saúde.

9.2. COMPARAÇÕES COM TRABALHOS SEMELHANTES

Estudos que usam técnicas de MD buscam encontrar conhecimento em diferentes áreas, o que também acontece na Saúde com os mais variados enfoques. Nesta seção, serão discutidos trabalhos relacionados ao tema deste trabalho.

Incialmente foi feita uma pesquisa por trabalhos com títulos semelhantes ao deste trabalho em três bases eletrônicas utilizadas no Brasil e América Latina, o LILACS (Literatura Latino-Americana e do Caribe em Ciências da Saúde), o SciELO (*Scientific Electronic Library Online*) e o Portal de Periódicos da Capes.

Na tabela 8 podem ser vistos os termos pesquisados, tantos em português, quanto em inglês, e os respectivos resultados com o número de trabalhos semelhantes. Contudo, o que mais aconteceu foi o retorno de referências indiretas as palavras chaves dos termos descritores inseridos nas buscas.

Tabela 10 - Pesquisas dos termos semelhantes

Termos de busca	LILACS	SciELO	Capes
Mineração de dados médicos em bases abertas	0	0	9
Medical data mining on open databases	0	0	25.000
Mineração de dados médicos	16	26	77
Medical data mining	36	2.500	58.000

Fonte: Autoria própria.

Os trabalhos encontrados que estão relacionados a bases de dados abertas geralmente focam em questões relacionadas a disponibilização destes dados nas mais diferentes esferas de governo. Com exemplos muito conhecidos, como o DATASUS (DATASUS, 2018) no caso da trabalhos da área de saúde ou portal o

Portal Brasileiro de Dados Abertos (DADOS ABERTOS, 2018) para trabalhos em geral.

Com frequência são encontrados trabalhos que comparam a disponibilização de dados abertos pelos municípios, como fizeram DOMINGUEZ e ALMADA (2018). Eles compararam seis capitais do Brasil em diversas questões relativas a áreas como Educação, Administração Pública, Infraestrutura, Mobilidade, Economia, Saúde, entre outras.

Os resultados mostram que não foi encontrado, em nenhum dos portais de dados abertos, algum conjunto de dados com a uma avaliação máxima nos quesitos técnicos. Os problemas mais comuns encontrados foram: dificuldade de trabalhar os dados, a indisponibilidade de download da base completa, conjunto de dados incompleto e indisponibilidade de formato aberto. Além disso, constatou-se que a maioria das bases de dados publicadas correspondem aos temas Administração Pública e Infraestrutura Urbana, com poucas dando destaque a área da saúde.

9.2.1. Síntese dos Estudos Analisados

Já os trabalhos encontrados com tema diretamente relacionados a MD e aqui comparados foram cinco. COSTA (2007) aplicou MD em duas bases de dados na área da saúde sobre o Câncer de Mama, identificando assim o número de pessoas com probabilidade significativa de desenvolver Câncer de Mama Benigno ou Câncer Maligno. Os algoritmos que melhor apresentaram resultados dentre os vários que foram utilizados foram os dos classificadores JRip, OneR e ZeroR baseados em Regras, e os J48, REPTree, LMT de Árvores de Decisão e o MultilayerPerceptron de Redes Neurais Artificiais.

Já GREGORY e PRETTO (2016), fizeram um estudo para auxiliar uma empresa de promoção à saúde na tomada de decisões em relação aos seus clientes. Dentre as tarefas de MD, foram utilizadas Associação e Clusterização. As informações necessárias para a MD foram extraídas do software de gestão de clientes da empresa analisada.

Os atributos que foram analisados são a região do paciente, gênero, nível de glicose, pressão arterial, tabagismo, consumidores de álcool, atividades físicas e alimentação. Mesmo com muitos problemas relacionados as informações extraídas

do sistema da empresa, pois os dados estavam muito incompletos ou não informados, foi possível descobrir as relações e correlações da saúde das pessoas dos Vales do Taquari e Rio Pardo.

MACIEL et al. (2015) descreveu, passo a passo, a aplicação de um processo de descoberta de conhecimento em banco de dados (KDD) no domínio da triagem médica. O foco do estudo foi nas fases de pré-processamento e na mineração de dados, especificamente na tarefa de classificação.

Foi feita a aplicação do algoritmo classificador J48 no WEKA, utilizando uma abordagem sensível a custo. A acurácia apresentada não teve um valor muito alto, de 59,33%. Contudo, isso não foi visto como um problema, pois a acurácia não foi considerada como o único critério de sucesso.

STEINER et al. (2006) mostraram a influência da análise exploratória dos dados no desempenho das técnicas de MD quanto à classificação de novos padrões por meio da sua aplicação a um problema médico. Além disso, eles compararam o desempenho das técnicas, visando obter a técnica com o maior percentual de acertos. Para o problema em estudo, Programação Linear e Redes Neurais foram as técnicas que apresentaram os menores percentuais de erros para os conjuntos de testes.

FRANÇA et al. (2016) aplicaram o KDD sobre Ordens de Serviço de informática de uma instituição hospitalar do Paraná, buscando identificar novas práticas a serem aplicadas na gestão. O objetivo foi identificar os padrões de relacionamento entre os diversos setores do hospital com a área de informática, naquilo que tangia os atendimentos a problemas de Tecnologia de Informação. Para os experimentos foi utilizada a tarefa de classificação, sendo avaliados três métodos, que possibilitaram a obtenção de regras com potencial de validação por especialista. O resultado demonstrou que o KDD obteve conhecimento relevante para auxílio à tomada de decisão com vistas ao investimento em melhores práticas para ganho de qualidade dos serviços de TI.

Nos trabalhos comparados, foram levantados aspectos citados pelos autores como importantes nas fases do KDD, os mesmos aspectos que foram tratados ao longo deste trabalho. Estes aspectos foram agrupados por similaridade, de acordo o problema descrito, os enfoques aplicados e as soluções encontradas. As informações foram dispostas na tabela 9, que apresenta nas linhas os aspectos

semelhantes e nas colunas os autores dos trabalhos. Utilizou-se uma abordagem semelhante à usada por CARVALHO et al. (2014).

Tabela 11 - Comparações com Trabalhos Semelhantes

Aspectos semelhantes encontrados	COSTA (2007)	GREGORY e PRETTO (2016)	MACIEL et al. (2015)	STEINER (2006)	FRANÇA et al. (2015)
Utilização do KDD	X	X	X	X	X
Avaliação da qualidade dos dados originais	-	X	X	X	X
Importação dos dados originais para um banco de dados	-	X	-	-	X
Construção de software específico para auxílio do processo	-	X	-	-	-
Uso de diferentes tipos de tarefas e algoritmos	X	X	-	X	X
Discussão da relevância dos padrões encontrados	-	X	X	X	X
Auxílio ao planejamento em saúde para médico/gestores	X	-	-	X	X

Fonte: Autoria própria.

10. CONCLUSÃO E TRABALHOS FUTUROS

Este capítulo apresenta as conclusões desta pesquisa, bem como suas contribuições e possíveis implicações. Sugere ainda os trabalhos futuros que poderiam ser realizados.

10.1. CONCLUSÕES

Este trabalho analisou a base de dados pública E-Saúde buscando a identificação de padrões (conhecimento) que poderiam auxiliar na tomada de decisão de gestores administrativos/médicos das Unidades Básicas de Saúde do sistema de Saúde da cidade de Curitiba, a partir das informações obtidas por meio de técnicas de mineração de dados.

A questão central investigada foi o quanto a base de dados pública E-Saúde, passando por um processo de descoberta de conhecimento (KDD) poderia fornecer informações úteis para auxiliar os tomadores de decisão no ambiente da saúde.

O KDD aplicado envolveu uma série de fases, que foram distribuídas em três principais etapas: pré-processamento (importação, limpeza e ajustes), mineração de dados (aplicação dos algoritmos) e pós-processamento (interpretação).

No pré-processamento foi realizada a importação dos dados disponibilizados no portal de dados abertos de Curitiba. Essa importação foi conduzida com sucesso para um banco de dados MySQL, por uma aplicação na linguagem PHP construída especialmente para esta finalidade. O processo de importação não foi trivial e levou mais de um mês até ser concluído.

Também foi necessário realizar trabalhos de limpeza de dados, como retirar caracteres especiais que paravam o processo de importação, além de adequações quando os atributos vazios lido retornavam valores nulos. Isso indica que a base E-Saúde, não está pronta para ser usada de forma direta por softwares mineradores de dados.

Antes da importação, havia desconhecimento de como estaria a base de dados. Não se sabia se ela tenderia a ser completa, ou seja, a presença da maioria das instâncias contendo informações, ou se tenderia a ser mais esparsa, com muitos atributos inexistentes.

Inspecionando os dados, constatou-se que a frequência dos dados nos 37 atributos disponibilizados, é alta para a maioria dos atributos, mostrando que a base E-Saúde possui uma característica favorável para a mineração de dados.

Constatou-se a presença de cinco pares de atributos com código e descrição, o que pode ser considerado como atributo repetidos para o processo de mineração de dados, pois apenas um deles já é suficiente. Isso indica a necessidade de adequar a E-Saúde para a MD.

Estudando os atributos fornecidos pela PMC na base de dados, foi possível a criação de 11 novos atributos. Isso foi feito com intuito de auxiliar as análises dos dados, esforço feito que melhorou a previsão dos algoritmos dos classificadores utilizados.

Com o total de 48 atributos, 37 originais, mais os 11 criados, foi necessário realizar uma seleção dos atributos que realmente tivessem relevância no fornecimento de informações significativas da base dados. Por isso, foi reduzida a quantidade para 26 atributos, entre os originais e os criados.

Na etapa da mineração dos dados, foi realizado uma análise nas unidades de saúde que compõem o Distrito Sanitário do Boqueirão e do Cajuru, referente a Resolutividade das Unidades de Saúde.

Para ambos os estudos, foi utilizada uma combinação dos algoritmos classificadores J48 (Árvore de Decisão) e JRip (Regras). Isso facilitaria a interpretação dos resultados por um gestor ou médico, pois associando a simplicidade de leitura do formato das Árvores de Decisão com o formato textual das Regras, não deverá ser necessário conhecimento em programação ou banco de dados.

Na análise da Resolutividade, foi possível evidenciar os atributos que contribuem para a resolução de um acesso do paciente a uma unidade de saúde. Na ordem de maior influência, foram os atributos “Indica se Ocorreu Solicitação de Exames”, “Indica se Houve Encaminhamento para Atendimento de Especialista” e “Indica se Desencadeou Internamento”. Ambos os classificadores apresentaram

uma boa precisão na classificação das instâncias, inclusive induzindo modelos com acurácia quase idênticos, próximo a 75%.

Por exemplo, uma regra que o classificador JRip evidenciou, foi que na unidade de saúde Menonitas (`descr_unidade=UMS_MENONITAS`), o conjunto de médicos Clínicos (`descr_CBO=MEDICO_CLINICO`) são os que mais fazem solicitações de exames (`solicit_exames=Sim`).

Desta forma seria possível a gestores, das Unidades Saúde ou mesmo do Distrito de Sanitário, fazerem acompanhamentos tanto dos recursos materiais como humanos, disponibilizando, treinamentos e reciclagens para os profissionais das Unidades de Saúde no sentido de solicitarem exames dentro dos protocolos existentes.

Assim, é possível concluir que conseguiu-se extrair informações úteis da base de dados E-Saúde que podem potencialmente suportar o processo de tomada de decisão de gestores.

Espera-se que esse trabalho contribua mostrando que mesmo tendo limitações, a base de dados E-Saúde tem seus méritos, e é viável para ser utilizada em mais estudos em outros trabalhos.

10.2. TRABALHOS FUTUROS

Trabalhos futuros para analisar as potencialidades da base de dados abertos E-Saúde utilizando as inferências fornecidas pela mineração de dados, podem mostrar aperfeiçoamentos, tanto procurando outras relações dentro da base quanto com sugestões para a PMC para acrescentar atributos, possibilitando mais e diferentes tipos análises.

Dentro das bases atuais, seria interessante estudar a dinâmica das especialidades não analisadas nesse trabalho, como dos profissionais de nível superior como, por exemplo, Psicólogos, Nutricionistas e Fisioterapeutas.

Também um estudo dos atendimentos Odontológicos poderia ser realizado. Contudo, a época do trabalho, a quantidade de dados era muito inferior à das outras bases, cerca de mil vezes menos.

Por fim, outra sugestão de um trabalho futuro seria entrar em contato com o ICI, empresa que faz o processamento de dados da PMC e solicitar a inclusão de mais informações (atributos) na base de dados a serem disponibilizadas no portal. Por exemplo, seria interessante se existisse a diferenciação entre os atendimentos relativos às UBS e às USF (caso seja possível) nas unidades que possuem as duas. Isso possibilitaria a análise dos dados específicos de cada tipo de serviço realizado na unidade.

Também, a base atual não disponibiliza nenhuma referência ao endereço do paciente, usuário do sistema de saúde, ainda que o anonimato faça-se necessário, principalmente por serem dados médicos. Se pelo menos o bairro do paciente fosse disponibilizado, um novo leque de opções se abriria para muitos tipos de pesquisas relacionadas a ocorrências e incidências de vários fatores em função dos 75 bairros da cidade.

11. REFERÊNCIAS

ALVES FILHO, Antônio e BORGES, Livia de O. **A Motivação dos Profissionais de Saúde das Unidades Básicas de Saúde** Psicol. cienc. prof. vol.34 n.4 2014 out-dez, Brasília, 2014.

AMARAL, Fernando **Aprenda Mineração de Dados: Teoria e Prática** ed. Alta Books, 2016.

ARFF **Attribute-Relation File Format** Disponível em <https://www.cs.waikato.ac.nz/ml/weka/arff.html> Acessado em 15/07/2018.

ATENÇÃO **Atenção Primária em Curitiba** Disponível em <http://www.saude.curitiba.pr.gov.br/atencao-basica/atecao-primaria.htmll> Acessado em 01/07/2018.

AVENTURIER, Pascal e ALENCAR, Maria C. de. **Os desafios de dados de pesquisa abertos em saúde** Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, v.10, n.3 p.1-19, jul.-set. Rio de Janeiro, 2016.

BENTO, Evaldo JUNIOR **Desenvolvimento web com PHP e MySQL** ed. Casa do Código, 2013.

BREIMAN L; FRIEDMAN J. H.; OLSHEN R. A.; STONE C. J. **Classification and regression trees.** Monterey: Wadsworth and Brooks; 1984.

CARDOSO, Sabrina B. de A.; PEREIRA, Maurício F.; PEREIRA, Juliana **Contribuição do Distrito Sanitário Sul na Gestão da Saúde Pública: análise de coordenadores de Centros de Saúde do Município de Florianópolis** Coleção Gestão da Saúde Pública - v.9 p.186-198, 2013.

CARTEIRA **Carteira de Serviços de Saúde do SUS de Curitiba** Disponível em <http://www.saude.curitiba.pr.gov.br/vigilancia/sanitaria/saude-do-trabalhador/13-geral/carta-do-sus-curitiba.html> Acessado em 03/07/2018.

CARVALHO, Deborah R.; MOSER, Auristela D.; SILVA, Verônica A. da; DALLAGASSA, Marcelo Rosano **Data Mining applied to physiotherapy** Fisioterapia e Movimento, v.25, n.3, p.595-605, jul.-set. Curitiba, 2012.

CARVALHO, Deborah R.; ESCOBAR, Leandro F. A.; TSUNODA, Denise F. **Pontos de Atenção para o Uso da Mineração de Dados da Saúde** Informação & Informação. v19, p249-273, Londrina, 2014.

COHEN, William W. **Fast Effective Rule Induction** Twelfth International Conference on Machine Learning p.115-123, 1995.

COSTA, Claudio N.; COUTINHO, Jonatas V.; MAGALHÃES, Lúcia H. de; ARBEX, Márcio A. **Descoberta de conhecimento em bases de dados** FESJ Revista Eletrônica, 2018. Disponível em <http://fsd.edu.br/revistaeletronica/arquivos/2Edicao/artigo9.pdf> Acessado em 08/07/2018.

COSTA, Rodrigo B. R. **da Aplicação do Processo de Mineração de Dados para Auxílio à Gestão do Pronto-Socorro de Clínica Médica do Hospital Universitário de Brasília Universidade Federal de Brasília** 2007. Disponível em <monografias.cic.unb.br/-dspace/bitstream/123456789/120/1/RODRIGO_BERNARDES_MONOGRAFIA.-pdf> Acesso em 01/03/2018.

DADOS **Base de Dados Abertos** Disponível em <http://www.curitiba.pr.gov.br/dadosabertos/consulta/?grupo=1> Acessado em 26/02/2018.

DADOS ABERTOS **Portal Brasileiro de Dados Abertos** Disponível em <http://dados.gov.br/pagina/sobre> Acessado em 08/07/2018.

DATASUS **Departamento de Informática do SUS** Disponível em <http://www2.datasus.gov.br/DATASUS/index.php?area=01> Acessado em 28/03/2018.

DATE, C. J. **Introdução a Sistemas de Bancos de Dados** 8ed Campus, 2004.

DEGANI, Vera C. **A resolutividade dos problemas de saúde: opinião de usuários em uma Unidade Básica de Saúde** (Dissertação) Mestrado em

Enfermagem da Escola Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

DISTRITOS **Distritos Sanitários de Curitiba** Disponível em <http://www.saude.curitiba.pr.gov.br/a-secretaria/localizacao-de-servicos-da-saude.html> Acessado em 03/07/2018.

DOMINGUEZ, Maria e ALMADA, Maria P. **Dados abertos governamentais: um estudo de caso dos portais de seis capitais brasileiras** 1 Congresso do Instituto Nacional de Ciência & Tecnologia em Democracia Digital Salvador, 2018.

E-SAÚDE **E-SAÚDE, Instituto das Cidades Inteligentes** Disponível em <http://www.ici.curitiba.org.br/conteudo/saude/51> Acessado em 08/07/2018.

ELMASRI, Ramez. e NAVATHE, Shamkant **Conceitos de Data Mining In: Sistemas de Banco de Dados** Pearson Addison Wesley, p.624-645, São Paulo, 2005.

ESF **Estratégia Saúde da Família (ESF)**. Disponível em <http://portalms.saude.gov.br/acoes-e-programas/saude-da-familia> Acessado em 01/07/2018.

FAYYAD, U.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic **From Data Mining to Knowledge Discovery in Databases** AI magazine, 17(3): p.37, Boston, 1996a.

FAYYAD U., USAMA M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic e UTHURUSAMY, Ramasamy **Advances in Knowledge Discovery and Data Mining** p.37-54, AAAI Press, The Mit Press, 1996b.

FERREIRA, Eduardo A. L. **Mineração de Dados Aplicada à Dados Médicos** Dissertação (Mestrado) Programa de Pós-Graduação Em Bioinformática 68f Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

FRANÇA, Gilson E. F.; CARVALHO, Deborah R.; TSUNODA, Denise F. **Descoberta de Padrões em Ordens de Serviço de Tecnologia da Informação em Hospital** Revista de Gestão em Sistemas de Saúde. v5 p41(11), 2016.

FREITAS A. A.; LAVINGTON S. H **Mining very large databases with parallel processing** Norwell Kluwer Academic Publishers, 1998.

GABARDO, A. C. **Análise de Redes Sociais - Uma Visão Computacional** Ed. Novatec, São Paulo, 2015.

GREGORY, Guilherme e PRETTO, Fabrício **Mineração de Dados para Descoberta de Conhecimento em Dados de Promoção à Saúde Revista Destaques Acadêmicos** Lajeado, v8, n4, 2016.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques** 3ed. Morgan Kaufmann Publishers Inc, San Francisco, USA, 2011.

INDA **Instrução Normativa da INDA** Disponível em <http://dados.gov.br/pagina/instrucao-normativa-da-inda> Acessado em 08/07/2018.

KHABAZA, Tom **Data Mining & Predictive Analytics,** 2010. Disponível em <http://www.khabaza.com/> Acessado em 10/07/2018.

LIBRELOTTO, Solange R. e MOZZAQUATRO, Patricia M. **Análise dos Algoritmos de Mineração J48 e Apriori Aplicados na Detecção de Indicadores da Qualidade de Vida e Saúde** Revista Interdisciplinar de Ensino Pesquisa e Extensão, v1, 2013.

MACIEL, Thales V.; SEUS, Vinicius R.; MACHADO, Karina dos S.; BORGES, Eduardo N. **Mineração de dados em triagem de risco de saúde** Revista Brasileira de Computação Aplicada, v7 n2, p26-40, Passo Fundo, 2015.

MONARD, Maria C. e PRATI, Ronaldo C. **Aprendizado de Máquina Simbólico para Mineração de Dados XIII** Escola Regional de Informática da SBC - Santa Catarina. 1 ed. Sociedade Brasileira de Computação, Florianópolis, 2005.

MySQL **Global Development Group. The world's most advanced open source database** Disponível em: <https://www.mysql.com/> Acessado em 20/02/2018.

NAKAMURA, Cristiane Y.; OTERO, Sandra D.; CARVALHO, Deborah R. **Data Mining For Coping Mother-To-Child Syphilis Transmission** XV Congresso Brasileiro de Informática em Saúde p.171, Goiânia, 2016.

NUNES, Altacílio A.; CACCIA-BAVA, Maria do C. G. G.; BISTAFA, Maria J.; PEREIRA, Luciana C. R.; WATANABE, Marlívia C.; SANTOS, Vânia e DOMINGOS, Nélio A. M. **Resolubilidade da estratégia saúde da família e unidades básicas de saúde tradicionais: contribuições do Pet-Saúde** Rev. bras. educ. med. vol.36 n.1 supl.1 jan.-mar., Rio de Janeiro, 2012.

OKF **Open Knowledge Brasil** Disponível em <https://opendefinition.org/od/1.1/pt/> Acessado em 08/07/2018.

OLIVEIRA, Saionara N.; RAMOS, Bianca J.; PIAZZA, Marina; PRADO, Marta L. do; REIBNITZ, Kenya S. e SOUZA, Cilonei A. **Unidade de Pronto Atendimento - UPA 24h: Percepção da Enfermagem** Texto Contexto Enfermagem, 2015 Jan-Mar 24(1): p.238-44, Florianópolis, 2015.

OLIVEIRA JÚNIOR, José G. de **Identificação de Padrões para a Análise da Evasão em Cursos de Graduação Usando Mineração de Dados Educacionais** 86 f. Dissertação (Mestrado) Programa de Pós-Graduação em Computação Aplicada Universidade Tecnológica Federal do Paraná, Curitiba, 2015.

OPEN GOVERNMENT **Open Government Data. 8 Principles of Open Government Data** Disponível em <http://www.opengovdata.org/home/8principles> Acessado em 08/07/2018.

PEREIRA, Tamara A. e SOUZA JUNIOR, Amauri H. de **Uma formulação para a máquina de aprendizagem mínima baseada em programação linear** XXXVII Congresso da Sociedade Brasileira de Computação p.2572, 2017.

PHP **Referência da Linguagem** Disponível em <http://php.net/manual/pt_BR/intro-whatis.php> Acessado em 15/07/2018.

PORTAL **Portal de Dados Abertos da Prefeitura Municipal de Curitiba** Disponível em <http://www.curitiba.pr.gov.br/dadosabertos/consulta/> Acessado em 26/02/2018.

QUINLAN, J. Ross **C4.5: Programs for Machine Learning** Morgan Kaufmann Publishers, San Mateo, CA, USA 1993.

REZENDE, Solange O. **Sistemas Inteligentes Fundamentos e Aplicações** ed. Manole, São Paulo, 2005.

ROMERO, C.; VENTURA, S.; ESPEJO, P. G.; e HERVÁS, C. **Data mining algorithms to classify students. In The First International Conference on Educational. International Conference on Educational Data Mining** p.8-17, Córdoba University, Spain, 2008.

SANTOS, Asaffe C. M. dos **Aprendizado de máquina aplicado ao diagnóstico de Dengue** SBC ENIAC, p. 697-708, Recife - PE, 2016.

SAÚDE **Histórico da Secretaria Municipal de Saúde de Curitiba** Disponível em <http://www.saude.curitiba.pr.gov.br/a-secretaria/historico-da-secretaria.html> Acessado em 28/03/2018.

SHIBA, Sonia K. **Desenvolvimento de Modelo de Processo de Extração de Conhecimento em Banco de Dados para Sistemas de Suporte à Decisão** (Dissertação) Programa de Pós-Graduação em Engenharia Elétrica, São Paulo, 2008.

SILVA, Clayton F. da; VAZ, Wesley; SANTOS, Erick M. F. dos; BALANIUK, Remis; CHAVES, Mônica C. **Dados abertos: uma estratégia para o aumento da transparência e modernização da gestão pública** Revista do TCU 131 2014 set-dez p.22-29, 2014.

SILVA, Inara A. F. **Descoberta de conhecimento em base de dados de monitoramento ambiental para avaliação da qualidade da água** 2007 Dissertação (Mestrado) Programa de Pós-Graduação em Física e Meio Ambiente, Universidade Federal de Mato Grosso, Cuiabá, 2007.

SILVA, V. J. da; GOMES, C. E. M.; SANTANA, S. S.; LUCENA JUNIOR, V. F. de **Sistema Inteligente para Gerenciamento De Medicação Em Ambientes Residenciais** XXV Congresso Brasileiro de Engenharia Biomédica p.2159, 2016.

SOARES JUNIOR, Jair S. e QUINTELLA, Rogério H. **Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica** RAP 39(5):p.1077-1107, set.-out., Rio de Janeiro, 2005.

SOCZEK, Felipe e ORLOVSKI, Regiane **Mineração de Dados: Conceitos e Aplicação de Algoritmos em Uma Base de Dados na Área da Saúde** Revista Científica Semana Acadêmica ano MMXIV n.50, Fortaleza, 2014.

SOUZA, Jackson W. da C. e FELIPPO, Ariani Di **Caracterização Da Complementaridade Temporal: Subsídios Para Sumarização Automática Multidocumento** v.62 n.1 p.125-150, São Paulo, 2018.

SOUZA, Primavera B. de **Uma estratégia baseada em algoritmos de mineração de dados para validar plano de operação de voo a partir de predições de estados dos satélites do INPE** (Tese) 173f INPE, São José dos Campos, 2011.

STEINER, Maria T. A.; SOMA, Nei Y.; SHIMIZU, Tamio; NIEVOLA, Júlio C.; NETO, Pedro J. S. **Abordagem de um Problema Médico por meio do Processo de KDD com ênfase à Análise Exploratória dos Dados** Revista GESTÃO & PRODUÇÃO, v13 n2, p325-337, 2006.

TAN, P. N.; STEINBACH, M. e KUMAR, V. **Introdução ao Data Mining - Mineração de Dados** Ed. Ciência Moderna, 2009.

TRONCHONI, Alex B.; PRETTO, Carlos O.; ROSA, Mauro A. da **Descoberta De Conhecimento Em Base De Dados De Eventos De Desligamentos De Empresas De Distribuição** Revista Controle & Automação v..21 n2 mar-abr, 2010.

UPA. **Unidade de Pronto Atendimento (UPA)** Disponível em <http://portalms.saude.gov.br/acoes-e-programas/upa/sobre-o-programa> Acessado em 01/07/2018.

VASCONCELOS, Benitz de S. **Mineração de Regras de Classificação com Sistemas de Banco de Dados Objeto-Relacional** (Dissertação) Programa de Pós-Graduação em Informática Campina Grande, 2002.

VIEIRA, Andrws A. **Uma abordagem para estimação prévia dos requisitos não funcionais em sistemas embarcados utilizando métricas de software.** 2015 104 f. (Dissertação) – Programa de Pós-Graduação em Computação. Universidade Federal Do Rio Grande Do Sul. Porto Alegre, 2015.

WEKA **Data Mining with Open Source Machine Learning Software in Java** Disponível em <http://www.cs.waikato.ac.nz/ml/weka/> Acessado em 05/02/2018.

WEKA EXPLORER **Data Mining with Open Source Machine Learning Software in Java** Disponível em <https://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html> Acessado em 15/07/2018.

WITTEN, I. H., FRANK, E. e HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques** Morgan Kaufmann Publishers Inc, 3ed, San Francisco CA USA, 2011.

12. APÊNDICE

Diagrama dos atributos do DS do Boqueirão para o Estudo de Caso da Resolutividade

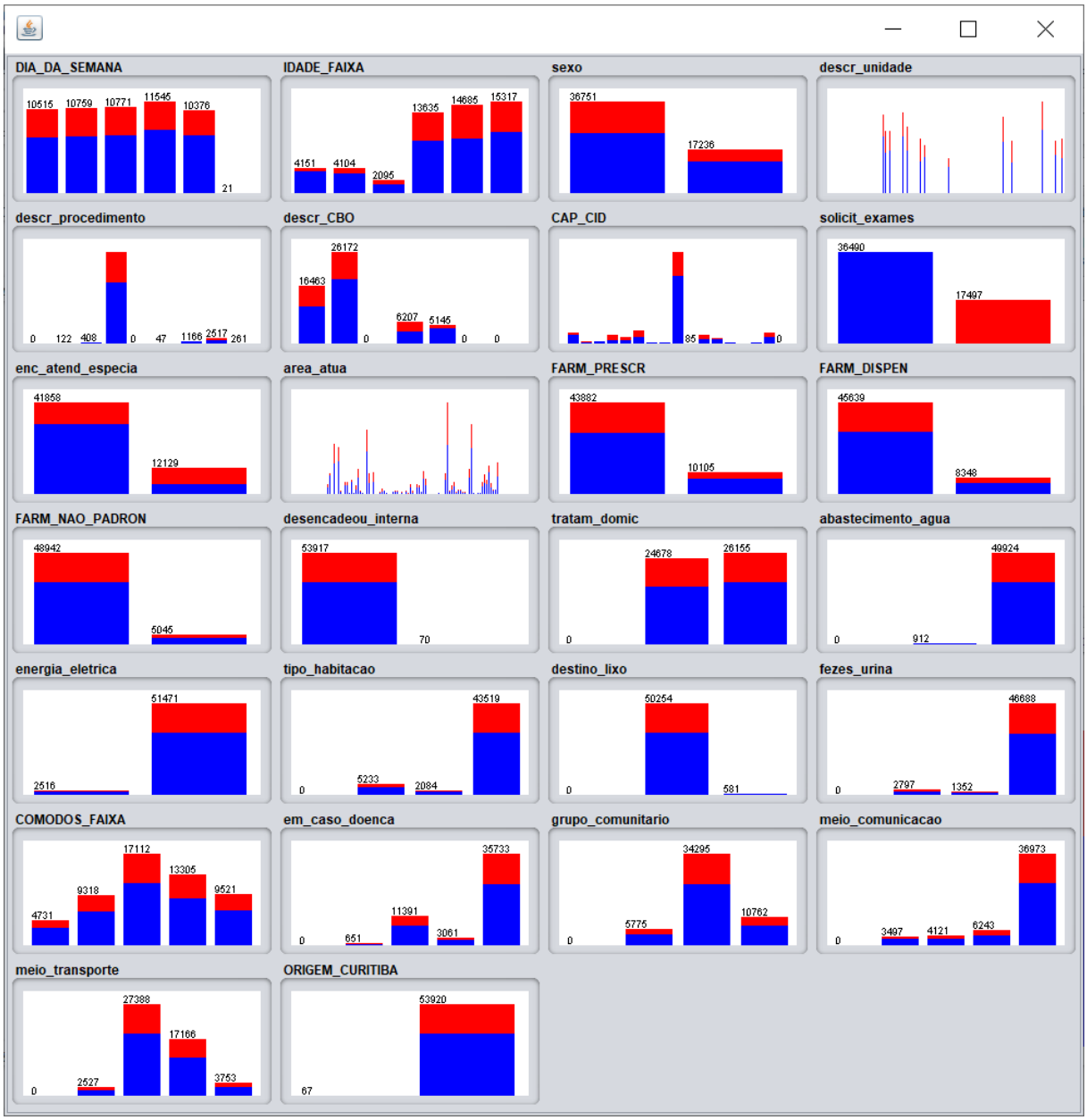
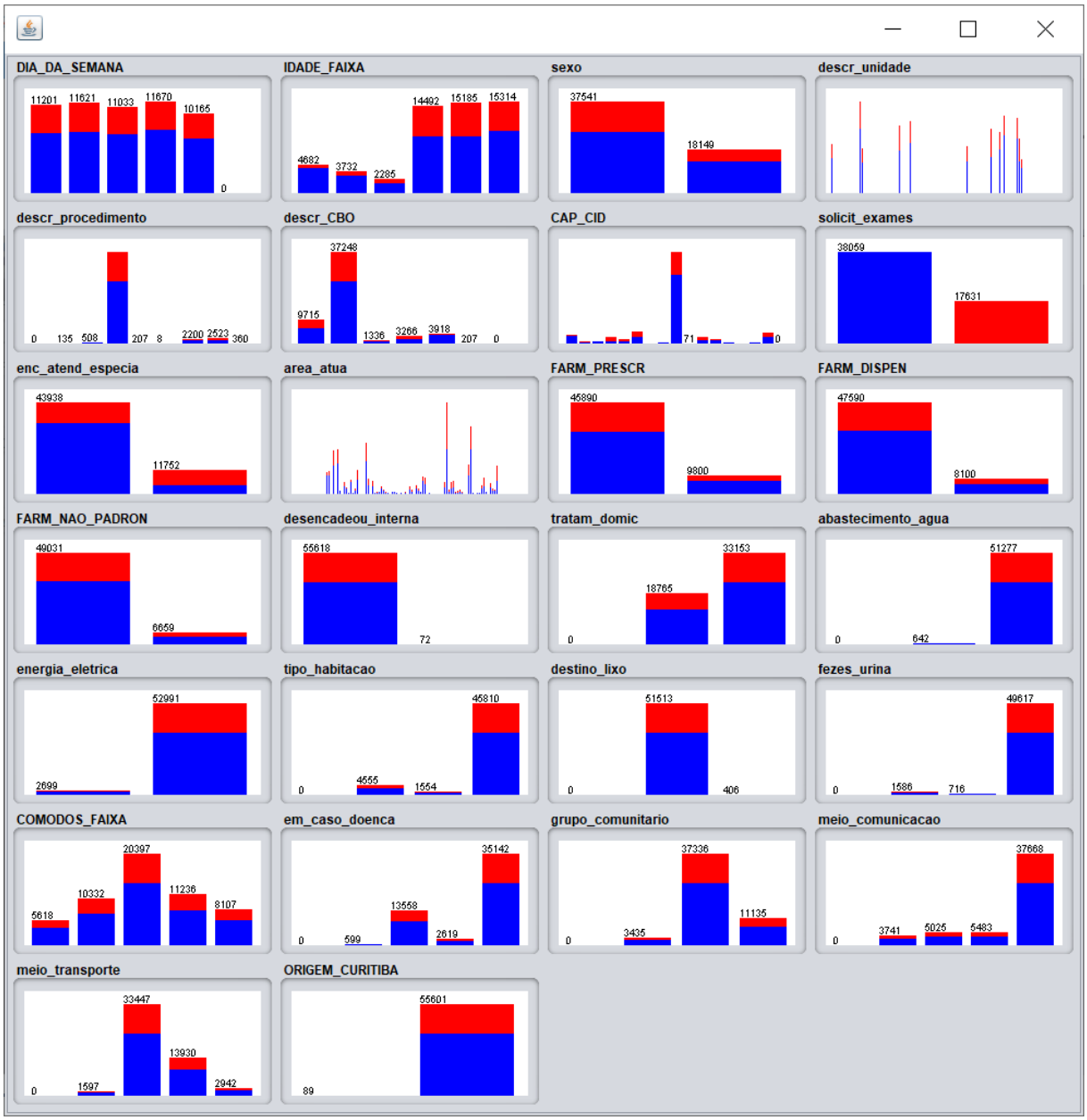


Diagrama dos atributos do DS do Cajuru para o Estudo de Caso da Resolutividade



13. ANEXOS

NAME
weka.classifiers.trees.J48

SYNOPSIS
Class for generating a pruned or unpruned C4.5 decision tree. For more information, see

Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

OPTIONS
seed -- The seed used for randomizing the data when reduced-error pruning is used.

unpruned -- Whether pruning is performed.

confidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning).

numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.

numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model.

batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.

reducedErrorPruning -- Whether reduced-error pruning is used instead of C.4.5 pruning.

useLaplace -- Whether counts at leaves are smoothed based on Laplace.

doNotMakeSplitPointActualValue -- If true, the split point is not relocated to an actual data value. This can yield substantial speed-ups for large datasets with numeric attributes.

debug -- If set to true, classifier may output additional info to the console.

subtreeRaising -- Whether to consider the subtree raising operation when pruning.

saveInstanceData -- Whether to save the training data for visualization.

binarySplits -- Whether to use binary splits on nominal attributes when building the trees.

doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).

minNumObj -- The minimum number of instances per leaf.

useMDLcorrection -- Whether MDL correction is used when finding splits on numeric attributes.

collapseTree -- Whether parts are removed that do not reduce training error.

CAPABILITIES
Class -- Missing class values, Binary class, Nominal class

Attributes -- Empty nominal attributes, Date attributes, Missing values, Binary attributes, Numeric attributes, Unary attributes, Nominal attributes

Additional
min # of instances: 0

NAME
weka.classifiers.rules.JRip

SYNOPSIS
This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.

The algorithm is briefly described as follows:

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one, DO:

1. Building stage:
Repeat 1.1 and 1.2 until the descrition length (DL) of the ruleset and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

1.1. Grow phase:
Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

1.2. Prune phase:
Incrementally prune each rule and allow the pruning of any final sequences of the antecedents;The pruning metric is (p-n)/(p+n) -- but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

2. Optimization stage:

after generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset.After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.
3. Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it. and add resultant ruleset to RS.
ENDDO

OPTIONS
seed -- The seed used for randomizing the data.

numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model.

batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.

folds -- Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules.

minNo -- The minimum total weight of the instances in a rule.

debug -- Whether debug information is output to the console.

optimizations -- The number of optimization runs.

doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).

checkErrorRate -- Whether check for error rate >= 1/2 is included in stopping criterion.

usePruning -- Whether pruning is performed.

CAPABILITIES
Class -- Missing class values, Binary class, Nominal class

Attributes -- Empty nominal attributes, Date attributes, Missing values, Binary attributes, Numeric attributes, Unary attributes, Nominal attributes

Additional
min # of instances: 3