UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS DE DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

LILIAN DE FÁTIMA PETROSKI

UMA ABORDAGEM DE DESCOBERTA DE CONHECIMENTO PARA O SUPORTE À GESTÃO MUNICIPAL DE SAÚDE

TRABALHO DE CONCLUSÃO DE CURSO

DOIS VIZINHOS
2021

LILIAN DE FÁTIMA PETROSKI

UMA ABORDAGEM DE DESCOBERTA DE CONHECIMENTO PARA O SUPORTE À GESTÃO MUNICIPAL DE SAÚDE

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Marcelo Teixeira

DOIS VIZINHOS
2021

LILIAN DE FÁTIMA PETROSKI

UMA ABORDAGEM DE DESCOBERTA DE CONHECIMENTO PARA O SUPORTE À GESTÃO MUNICIPAL DE SAÚDE

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de Aprovação: 16/dezembro/2021

Marcelo Teixeira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

Anderson Chaves Carniel
Doutorado
Universidade Federal de São Carlos

Jefferson Tales Oliva
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

DOIS VIZINHOS
2021

Dedico este trabalho ao meu esposo Samarone
e à minha mãe Noili.

AGRADECIMENTOS

Primeiramente agradeço a Deus, por iluminar meu caminho para que chegasse ao fim desse desafio. Durante esta pós-graduação muitas experiências foram vividas, aprendi a recomeçar, superar e não desistir.

Agradeço, especialmente, ao meu esposo Samarone e a minha mãe Noili, que sempre me apoiaram e incentivaram-me a vencer mais essa etapa da vida. Esta conquista tem o sabor das dificuldades superadas.

Agradeço aos professores que compartilharam seus conhecimentos ao longo do curso e especialmente ao professor Dr. Marcelo Teixeira pela paciência, compreensão e orientação no desenvolvimento deste trabalho.

> De tudo ficarão três coisas: a certeza de que estamos começando, a certeza de que é preciso continuar e a certeza de que podemos ser interrompidos antes de terminar. Fazer da interrupção um caminho novo. Fazer da queda um passo de dança. Do medo, uma escada. Do sonho, uma ponte. Da procura, um encontro. (Fernando Sabino)

RESUMO

Petroski, Lilian de Fátima. Uma abordagem de descoberta de conhecimento para o suporte à gestão municipal de saúde. 2021. 39 f. Trabalho de Conclusão de Curso – Curso de Especialização em Ciência de Dados, Universidade Tecnológica Federal do Paraná. Dois Vizinhos, 2021.

As ferramentas computacionais de suporte às políticas públicas, que permitem a população e aos gestores avaliarem a eficiência e eficácia dos gastos são cada vez mais relevantes. Na área da saúde milhares de atendimentos são prestados diariamente gerando um grande volume de dados sendo necessário o auxílio de ferramentas adequadas para obter conhecimento dos dados. Um meio para extrair conhecimento dessa imensidão de dados é através do KDD que permite a identificação de novas informações úteis, válidas e compreensíveis. Na área da saúde pública o KDD pode identificar informações úteis para suporte à gestão municipal de saúde. Este trabalho analisa a base do E-saúde do perfil de atendimento médico nas unidades municipais de saúde de Curitiba através da aplicação do KDD, com a ferramenta WEKA e o algoritmo Apriori na etapa de mineração de dados, para identificar informações sobre o atendimento e perfil dos pacientes para auxiliar à gestão. Os resultados obtidos são promissores com as 34 regras encontradas na mineração, que trazem informações sobre os atendimentos, especialidades dos médicos e distribuição dos atendimentos nas unidades de acordo com a faixa etária dos pacientes.
Palavras-chave: descoberta de conhecimento, KDD, mineração de dados, saúde.

ABSTRACT

Petroski, Lilian de Fátima. A cross data based approach to support municipal health management. 2021. 39 f. Trabalho de Conclusão de Curso – Curso de Especialização em Ciência de Dados, Universidade Tecnológica Federal do Paraná. Dois Vizinhos, 2021.

The computational tools to support public policies, which allow the population and managers to assess the efficiency and effectiveness of expenditures, are increasingly relevant. In the healthcare area, thousands of assistances are provided daily, generating a large volume of data, requiring the help of appropriate tools to obtain knowledge of the data. One way to extract knowledge from this vast amount of data is through KDD which allows the identification of new useful, valid and understandable information. In the area of public health, the KDD can identify useful information to support municipal health management. This work analyzes the E-health base of the medical care profile in the municipal health units of Curitiba through the application of KDD, with WEKA tool and the Apriori algorithm in the data mining stage, to identify information about the care and profile of patients to assist with management. The results obtained are promising with the 34 rules found in mining, which provide information on care, specialties of doctors and distribution of care in the units according to the age group of patients.

Keywords: knowledge discovery, KDD, data mining, health.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARRF	Attribute Relation File Format
CBO	Classificação Brasileira de Ocupações
CID	Classificação Internacional de Doenças
COD	Código
CSV	Comma Separated Values
DS	Distrito Sanitário
GB	Gigabyte
KDD	Knowledge Discovery in Databases
MB	Megabyte
QTDE	Quantidade
SGBD	Sistema Gerenciador de Banco de dados
SQL	Structured Query Language
UPA	Unidade de Pronto Atendimento
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

As ferramentas voltadas às políticas públicas que permitem a população e aos gestores avaliarem a eficiência e eficácia dos serviços são cada vez mais relevantes para a sociedade. Atualmente, percebe-se um grande volume de dados sendo gerados e armazenados em sistemas de gerenciamentos de banco de dados e com esse crescimento, reduz-se a possibilidade das pessoas entendê-los sem ferramentas adequadas (WITTEN; FRANK; HALL, 2011). Muitas vezes, os sistemas são utilizados como tabuladores de informação, e com a limitação orçamentária da gestão pública é sempre importante a atenção à eficiência e eficácia no uso de recursos (ARAÚJO; OLIVEIRA; SILVA, 2007). Nesse sentido, soluções computacionais são essenciais para automatizar a pesquisa de padrões e a descoberta de informações úteis em dados (WITTEN; FRANK; HALL, 2011).

No Brasil, a publicação de informações pelo governo é um dever conforme o princípio da publicidade previsto no art. 37 da Constituição Federal do Brasil. No entanto, apenas com o advento da *Lei de Acesso a Informação* (LAI), os governos das esferas federal, estadual e municipal passaram a ter a obrigação de publicar dados na internet em formato aberto e que possam ser acessados por cidadãos e processados por máquina (FEDERAL, 2021b). Desse modo surgiram os portais de dados abertos que podem ser acessados por qualquer pessoa nos sites das diversas entidades que compõe a administração pública.

Um dos exemplos de cumprimento das diretrizes da LAI é o portal de dados abertos de Curitiba, que concentra diversas bases com informações do município, sendo disponibilizadas como serviço público municipal, não havendo restrições quanto ao uso, finalidade ou atividade no limite das restrições legais e demais regulamentações. Uma das bases disponíveis no portal de dados abertos de Curitiba é a do sistema E-saúde, com dados dos atendimentos da rede de atenção à saúde que é composta por Unidades Básicas de Saúde, Unidades de Pronto Atendimento, Centros de Especialidades Médicas e Odontológicas, entre outros (CURITIBA, 2021b).

A partir desse exemplo do município de Curitiba é evidente que com o crescente volume de dados disponíveis, precisamos de meios para extrair informações e conhecimento dessa massa de dados. A base de dados do E-saúde do perfil de atendimento médico nas unidades municipais de saúde possui um grande volume de dados, somente no primeiro trimestre de 2020 são quase 900 mil atendimentos. Uma forma de extrair conhecimento desse tipo de dado é com a aplicação do processo de descoberta de conhecimento em base de dados (em inglês *Knowledge Discovery in Databases* (KDD)) (MOHAMED; LTIFI; AYED, 2013; ESCOBAR et al., 2019; HIROTA; PEDRYCZ, 1999).

O KDD se refere ao processo de descoberta de conhecimento em base de dados, sendo uma forma não usual de identificar padrões novos, válidos, compreensíveis e permite a exploração de uma grande quantidade de dados. O KDD passa por cinco etapas: seleção, pré-

processamento, transformação, mineração de dados e interpretação dos resultados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). No centro do processo de KDD está a etapa de mineração de dados, na qual é realizada a busca por informações ocultas, é uma tecnologia poderosa com grande potencial para ajudar as organizações a se concentrarem nas informações importantes para tomar decisões proativas baseadas em dados (RAMZAN; AHMAD, 2014; ESCOBAR et al., 2019).

Existem diversas áreas em que o KDD pode ser aplicado, como saúde, educação, indústria, finanças, marketing e detecção de fraudes. Na área da saúde a aplicação do KDD pode ser voltada para efetividade do tratamento, predição de diagnósticos, mapeamento de causas de doenças e de mortalidade, diagnósticos fisioterapêuticos a fim de auxiliar os profissionais da saúde nas questões cotidianas (SANTOS et al., 2016).

Na literatura encontramos trabalhos que aplicam o KDD em dados da saúde. Por exemplo, Escobar et al. (2019), Gregory e Pretto (2016), Paula e Prado (2012) utilizam o método na construção de modelos para identificar oportunidades para redução de custos em operadoras de plano de saúde. Outra abordagem é apresentada em Silva (2019), Feuser (2017) onde o KDD é aplicado em uma base de dados de atendimentos municipais prestados pelo *Sistema Único de Saúde* (SUS) para analisar as possíveis causas e correlações de doenças, enquanto Maciel et al. (2015) tem como objetivo usar o método para auxiliar na triagem dos pacientes atendidos em *Unidades de Pronto Atendimento* (UPAs) do SUS e Santos (2018) utiliza o método para análise da efetividade dos atendimentos e quanto a solicitação de exames em uma base municipal. Na área pública, a extração de conhecimento dos dados pode gerar informações que auxiliem não só os profissionais da saúde que atuam diariamente com os pacientes como também fornecer suporte aos gestores públicos, contribuindo para melhor destinação dos recursos e criação de ações preventivas, proporcionando mais qualidade no atendimento à população.

Este trabalho aplica o método do KDD na base de dados do E-saúde referente ao perfil de atendimento médico nas unidades municipais de saúde de Curitiba. Na etapa de pré-processamento dos dados é usado pipeline do python, para o armazenamento dos dados utiliza o *Sistema de Gerenciamento de Banco de Dados* (SGBD) PostgresSQL, e na mineração de dados a ferramenta *Waikato Environment for Knowledge Analysis* (WEKA) com aplicação do algoritmo Apriori para obtenção das regras de associação e posterior interpretação dos resultados.

Os próximos capítulos deste trabalho estão organizados da seguinte forma, no capítulo 2 será apresentada a revisão da literatura, o capítulo 3 aborda o referencial teórico e no capítulo 4 são apresentados as etapas da aplicação do KDD e os resultados. Finalmente o capítulo 5 conclui o trabalho e apresenta sugestões de trabalhos futuros e as limitações encontradas.

2 ESTADO DA ARTE

Este capítulo apresenta trabalhos existentes na literatura que aplicam KDD ou mineração em bases de dados da área da saúde. Para isto foi realizada pesquisa utilizando o Google acadêmico, pois esta plataforma concentra variadas fontes de artigos e pesquisas, as palavras-chaves para busca foram KDD, mineração de dados e saúde.

Na proposta de Silva (2019), Feuser (2017) as etapas do KDD são aplicadas em uma base de dados de prontuários eletrônicos do SUS. Ambos utilizam a ferramenta WEKA para descobrir informações correlacionadas nos prontuários e obter conhecimento útil dos dados, a fim de associar as doenças com as possíveis causas para auxiliar na prevenção de novos casos. Em Silva (2019) para a classificação aplica os algoritmos $C4.5$, $Bagging$ e $Boosting$, enquanto Feuser (2017) utiliza o algoritmo Apriori, ambos com a finalidade de obter informações que permitam a promoção de ações preventivas, facilitando e contribuindo para as análises dos profissionais de saúde.

Maciel et al. (2015) utilizam o processo de descoberta de conhecimento em um conjunto de dados do Sistema Único de Saúde (SUS) referente à triagem de risco de vida em uma UPA obtendo um modelo de classificação que auxilie os profissionais nessa atividade, para isso aplicam o algoritmo C4.5 no software WEKA, o modelo gerado apresentou uma acurácia de 59,33% e quando utilizado o aprendizado sensível a custo a acurácia diminuiu para 56,56%.

Em Escobar et al. (2019), Gregory e Pretto (2016), Paula e Prado (2012) os autores propõe modelos para identificar padrões em bases de dados de operadoras de plano de saúde, visando apoiar o planejamento, identificar pontos para ações preventivas de promoção à saúde para reduzir gastos futuros. Escobar et al. (2019), Gregory e Pretto (2016) utilizaram o algoritmo Apriori para as regras de associação, em Escobar et al. (2019) no pós-processamento dos padrões sequenciais é utilizada a ferramenta $Chrono\_Assoc$ e para as janelas de tempo o $Assoctemp$, enquanto Gregory e Pretto (2016) e Paula e Prado (2012) utilizam a ferramenta WEKA.

Santos (2018) aplica o KDD na base de dados do E-saúde do perfil de atendimentos médico e de atendimento de enfermagem do município de Curitiba do primeiro trimestre de 2017, utilizou o WEKA com os algoritmos classificadores J48 e JRip utilizando o método de $cross - validation$ com 10 folds, com objetivo de analisar a resolutividade dos atendimentos, ambos os algoritmos apresentaram acurácia próxima a 75%. Porém, a aplicação da mineração de dados foi em um conjunto de dados de dois distritos sanitários, o do Boqueirão e do Cajuru, classificando o atributo de solicitação de exames que é uma das causas de retorno dos pacientes para nova consulta.

Dentre os trabalhos analisados Silva (2019), Feuser (2017) e Maciel et al. (2015) a aplicação do KDD nas bases de dados da área de saúde pública são para analisar as possíveis causas de doenças, facilitar a triagem de pacientes a fim de auxiliar os profissionais da saúde

em seus diagnósticos. Não houve exploração das bases para obter informações que auxiliem a gestão. Santos (2018) utiliza o KDD para análise da efetividade dos atendimentos e quanto a solicitação de exames, proporcionando informações para a gestão do quão eficiente estão sendo os atendimentos. No entanto, utiliza um algoritmo de classificação e explora três atributos da base utilizada, aplicando o modelo em dados de apenas dois distritos sanitários de Curitiba.

Neste trabalho é aplicado o algoritmo Apriori, com a ferramenta WEKA para análise dos atendimentos que foram realizados nas unidades de saúde do município de Curitiba, na busca de padrões relacionados à especialização do médico, perfil dos pacientes e de zoneamento para contribuir com a análise dos recursos e para seu melhor gerenciamento.

3 REFERENCIAL TEÓRICO

Atualmente muitos dados são armazenados e as técnicas e ferramentas que buscam transformar estes dados em conhecimento propriamente dito são cada vez mais necessárias. Uma das técnicas utilizadas na extração de informações dos dados é a Descoberta de Conhecimento em Bases de Dados (em inglês Knowledge Discovery in Databases - KDD). O processo de KDD oferece inúmeras oportunidades para as organizações obterem valor dos dados (GRADY, 2016; ABDRABO et al., 2016; SANTOS et al., 2016).

O KDD busca dar sentido aos dados através de mapeamento de um grande volume de dados, é um processo não trivial de identificação de novos padrões que sejam válidos, úteis e compreensíveis, concentra-se no processo geral de descoberta de conhecimento a partir dos dados. Incluí como os dados são armazenados e acessados, como algoritmos podem ser escalados para conjuntos de dados massivos e ainda funcionar de forma eficiente, como os resultados podem ser interpretados e visualizados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

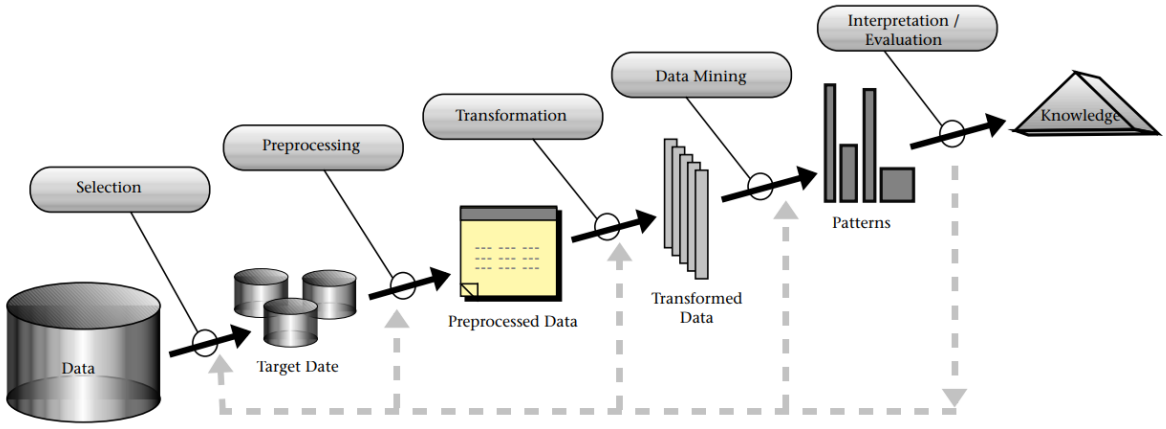
A técnica do KDD é o processo de extrair e refinar o conhecimento útil de bancos de dados, atualmente a sociedade é rica em dados e pode obter mais conhecimento desses dados(CERQUIDES; MANTARAS, 1997; HIROTA; PEDRYCZ, 1999).

3.1 Etapas do KDD

O KDD inclui uma mistura de técnicas de análise, como reconhecimento de padrões, agrupamentos, associações e análise visual, visando obter novos conhecimentos, mais profundos e com novas informações válidas e úteis (ABDRABO et al., 2016). Esse processo inclui cinco etapas contínuas e sequenciais com relação aos dados: seleção; pré-processamento e limpeza; transformação; mineração dos dados (MD) e interpretação dos resultados. A etapa da mineração de dados é a que visa propriamente descobrir os padrões nos dados (SANTOS et al., 2016; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Conforme Fig. 1 a primeira etapa é a seleção da base de dados que será estudada. Em seguida é realizado o pré-processamento etapa em que os dados podem ser filtrados, registros com problemas são removidos, valores padrões para dados nulos ou inválidos são atribuídos. Na etapa de transformação é realizada a formatação dos dados para que seja possível a realização da próxima etapa que é a mineração de dados onde os algoritmos são aplicados para a descoberta de padrões nos dados. E finalmente a etapa de interpretação da informação obtida dos dados, durante a aplicação do método é possível voltar as etapas anteriores. Após todas as etapas, pode ser concluído o processo de descoberta de conhecimento dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; SILVA, 2019).

Figura 1 – Etapas do KDD



Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

3.1.1 Seleção dos dados

A seleção é a primeira etapa no KDD, nessa fase é escolhida a base de dados pertencente a um sistema e que contém todas as variáveis e registros que serão analisados. O processo de seleção é complexo visto que os dados podem ter origem em fontes diferentes e estar em formatos diferentes e nesse caso é necessário realizar a conversão e integração dos dados (SILVA, 2019).

Consiste na criação ou seleção do conjunto de dados ou um subconjunto de variáveis ou ainda uma amostra de dados em que a descoberta de conhecimento será aplicada (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

3.1.2 Pré-processamento ou limpeza dos dados

As bases de dados podem apresentar inconsistências como registros incompletos, valores nulos e dados inválidos. O pré-processamento ou a limpeza dos dados é a segunda etapa do processo de KDD e uma das mais importantes, onde são realizadas análises preliminares e os tratamentos para eliminar dados inconsistentes, inválidos e redundantes. Nesta etapa podem ser aplicados filtros, atribuídos valores padrões para dados vazios, até a aplicação de técnicas de agrupamento para auxiliar na descoberta dos melhores valores para substituição dos ausentes, visando deixar a base o mais consistente possível para que seja realizada a próxima etapa (SILVA, 2019; CAMILO; SILVA, 2009).

A etapa de limpeza dos dados visa eliminar as inconsistências de modo que eles não influenciem no resultado dos algoritmos usados. É necessário fazer uma análise exploratória para obter conhecimentos sobre os dados analisando a consistência e mapeando os ruídos

que podem comprometer a qualidade do resultado, coletando informações e definindo as estratégias para enfrentar os campos com dados ausentes (CAMILO; SILVA, 2009; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Normalmente esta fase é a que consome mais tempo no processo de KDD (WITTEN; FRANK; HALL, 2011).

3.1.3 Transformação

A transformação dos dados consiste em selecionar os atributos, alterar o formato dos dados, transformar de classe para binário, aplicar métodos de redução de dimensionalidade que devem ser escolhidos de acordo com o objetivo da tarefa. Os métodos de redução ou transformação são utilizados para diminuir a quantidade de variáveis envolvidas no processo para melhorar o desempenho do algoritmo de análise (SILVA, 2019; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Nessa etapa deve ser analisado o tipo dos dados, que podem ser divididos em dois grupos: os quantitativos e os qualitativos. Os dados qualitativos contêm os valores nominais que também são chamados de categóricos. Já os dados quantitativos são representados usando números reais e podem ser discretos ou contínuos (CAMILO; SILVA, 2009; SANTOS, 2018). Antes de aplicar os algoritmos de mineração de dados é necessário explorar a base de dados para conhecê-los e para escolher os métodos adequados para a aplicação da mineração (CAMILO; SILVA, 2009).

Na etapa de transformação dos dados deve ser observado o tipo dos dados, pois alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. E por isso pode ser necessário transformar os valores numéricos em categóricos ou os categóricos em valores numéricos. Algumas técnicas podem ser utilizadas como suavização (remove valores errados), generalização (converter para valores mais genéricos), normalização (colocar as variáveis em uma mesma escala) e a criação de novos atributos a partir de outros já existentes (CAMILO; SILVA, 2009).

3.1.4 Mineração de dados

A Mineração de Dados é uma das técnicas mais promissoras da atualidade, devido a enorme quantidade de dados coletadas e armazenadas pelas companhias privadas e, no entanto, nenhuma informação útil é extraída e identificada. É aplicada em diversas áreas como por exemplo: mercado financeiro, transporte, telecomunicação, vendas, e além da iniciativa privada, o setor público também pode se beneficiar com a Mineração de Dados (CAMILO; SILVA, 2009).

O KDD se refere ao processo geral de descoberta de conhecimento útil a partir de dados, e mineração de dados se refere a uma etapa específica nesse processo. A mineração de dados é onde são aplicados os algoritmos específicos, por exemplo de classificação, regressão, agrupamento e associação, para extrair padrões dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A etapa de mineração de dados é a que recebe maior destaque na literatura (WITTEN; FRANK; HALL, 2011; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), tem como objetivo construir hipóteses, modelos, grupos e para isso existem diversos algoritmos como, por exemplo, o C4.5 e Apriori (SILVA, 2019).

De acordo com o problema a ser tratado será escolhida a tarefa de mineração de dados a ser aplicada (SANTOS et al., 2016). As tarefas podem ser de (SANTOS, 2018):

a - classificação: encontra padrões a partir dos atributos apresentados identificando tendências para aplicar em novos registros.

b - regressão ou estimativa: utilizada para valores númericos pode-se estimar o valor de uma variável analisando-se as demais.

c - associação: identifica os atributos que estão relacionados, obtendo regras.

d - clusterização: é o agrupamentos de atributos que possuem similaridade obtendo um subconjunto.

e - sumarização: descreve de maneira compacta um determinado conjunto de atributos com suas principais características.

O processo de mineração de dados não ocorre de forma totalmente automática e apesar de encontrarmos diversas ferramentas que auxiliam na sua execução, os resultados precisam de análise humana. No entanto, a mineração contribui de forma significativa no processo de descoberta de conhecimento, permitindo aos especialistas concentrarem esforços apenas em partes mais significativa dos dados (SANTOS et al., 2016).

3.1.4.1 WEKA

O WEKA surgiu em 1992 na Universidade de Waikato na Nova Zelândia devido a necessidade de um sistema que permitisse fácil acesso a algoritmos de mineração de dados e aprendizado de máquina, é desenvolvido em Java aproveitando-se das características de orientação a objeto (CARDOSO; SANTANA, 2019; WITTEN; FRANK; HALL, 2011).

Sendo ainda uma ferramenta bastante utilizada em pesquisas, por ser um sistema de código aberto e fornecer funcionalidades para pré-processamento, classificação, regressão, agrupamento, regras de associação e recursos para visualização dos dados. Os algoritmos podem ser executados na própria ferramenta proporcionando facilidade e agilidade na comparação do resultado de várias técnicas, e ainda pode ser incluída como parte de outros programas utilizando o seu processamento sem a sua interface gráfica (CAMILO; SILVA, 2009; WITTEN; FRANK; HALL, 2011).

O WEKA permite trabalhar com diferentes entradas de dados que podem ser desde um único arquivo no formato de relação de atributo (ARFF), arquivo separada por ponto e vírgula (CSV) ou através de conexão direta com o banco de dados (WITTEN; FRANK; HALL, 2011).

3.1.4.2 Regras de associação

As regras de associação têm como objetivo salientar relações não aparentes nos conjuntos de dados, contribuindo para a tomada de decisão (FEUSER, 2017). Trata-se de uma das técnicas mais conhecidas em mineração de dados para identificar o relacionamento de itens mais frequentes em um conjunto de dados, obtendo um resultado expresso na forma de X -> Y. É muito utilizado em problemas de cesta de compras (CAMILO; SILVA, 2009).

> O suporte mínimo (minsup) é a fração das transações que satisfaz a união dos itens do consequente com os do antecedente, de forma que estejam presentes em pelo menos s% das transações no banco de dados. A confiança mínima (minconf) garante que ao menos c% das transações que satisfaçam o antecedente das regras também satisfaçam o consequente das regras. (ROMÃO et al., 1999, p. 4).

Para uma regra de associação ser considerada forte ela deve atender a um suporte e confiança mínimos. O suporte corresponde à frequência em que os padrões ocorrem na base, e a confiança é o percentual dos registros que atendem a regra. Suporte e confiança não devem ser confundidos. A confiança é uma medida de qualidade da regra e o suporte corresponde a estatística. Normalmente as restrições de suporte se devem ao fato de que normalmente o interesse é nas regras com maior ocorrência, quando trata-se de motivos comerciais (AGRAWAL; IMIELIŃSKI; SWAMI, 1993). No entanto, um suporte mínimo muito elevado pode impedir a localização de regras menos frequentes.

Outra medida adotada para identificar regras interessantes é o $lift$, que representa qual a vantagem que uma regra oferece, pois estima quão mais frequente Y se torna quando X ocorre. É calculado dividindo a confiança pelo suporte (BAYARDO; AGRAWAL; GUNOPULOS, 2000; SILVA, 2004).

Um dos algoritmos mais utilizados para regras de associação é o Apriori (CAMILO; SILVA, 2009). O algoritmo Apriori pode trabalhar com um número grande de atributos, gerando a combinação entre eles, realiza buscas sucessivas em toda a base de dados, mantendo um ótimo desempenho em termos de tempo de processamento. O Apriori encontra na base de dados os chamados $itemsets frequentes$, fazendo uso de duas funções a $Apriori\_gen$ e a função $Genrules$. A primeira gera os conjuntos de itens candidatos, verificando o suporte mínimo determinado, e a segunda extrai as regras de associação considerando a confiança mínima estabelecida pelo usuário (AGRAWAL; SRIKANT et al., 1994; ROMÃO et al., 1999).

3.1.5 Interpretação

A última etapa do KDD é a interpretação ou avaliação do resultado da mineração de dados, onde é realizada a análise e seleção dos padrões relevantes para o problema pesquisado (SANTOS, 2018).

Após a descoberta dos padrões, o usuário precisa compreender de forma clara e objetiva e visualizar as informações descobertas e nesta etapa é importante a participação de especialista

no problema em estudo (PAULA; PRADO, 2012; COLLAZOS; BARRETO; PELLEGRINI, 2000).

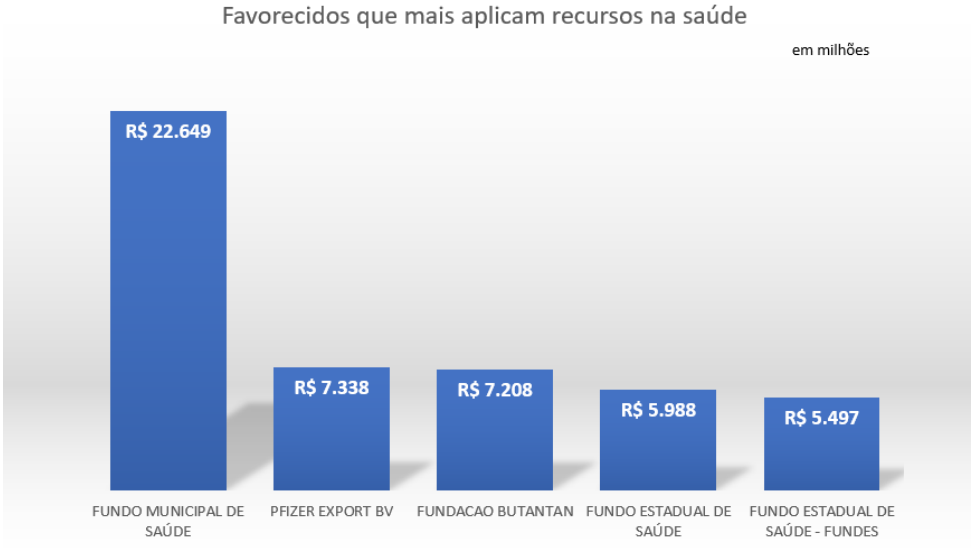
A interpretação dos resultados também envolve visualização dos padrões ou dos dados extraídos pelo modelo. Vale ressaltar que a qualquer momento é possível retornar as etapas anteriores para mais iterações com novas combinações e reconfigurações dos parâmetros buscando melhores resultados e conhecimento útil (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; COLLAZOS; BARRETO; PELLEGRINI, 2000).

4 UMA PROPOSTA PARA A MINERAÇÃO DO SISTEMA E-SAÚDE

Apresentados os conceitos do KDD e suas etapas, este capítulo tem o propósito de aplicá-los na base de dados do sistema E-saúde da prefeitura de Curitiba do perfil de atendimento médico nas unidades municipais de saúde, a fim de extrair conhecimento relacionado aos atendimentos, especialização do médico, perfil dos pacientes e zoneamento.

Há ainda outras bases de dados relacionadas a saúde disponíveis nos portais de dados abertos, tais como dados sobre ocupação hospitalar por Covid-19 (FEDERAL, 2021c) e da campanha de vacinação contra Covid-19 (FEDERAL, 2021a). Na literatura há alguns exemplos como Silva (2019) e Feuser (2017) que utilizam uma base de dados de prontuários eletrônicos do SUS e Maciel et al. (2015) que utilizam um conjunto de dados de triagem do Sistema Único de Saúde (SUS). Também há trabalhos com dados da saúde suplementar como Escobar et al. (2019), Gregory e Pretto (2016) e Paula e Prado (2012) com bases de dados de operadoras de plano de saúde. As pesquisas procuram descobrir conhecimento que auxilie os profissionais de saúde que estão na linha de frente no dia a dia ou ainda para auxiliar a iniciativa privada. Este trabalho tem como objetivo aplicar o KDD para contribuir com a gestão pública de saúde, em particular a gestão municipal, que planeja, executa, controla e avalia as ações dos serviços de saúde e ainda são responsáveis pela gestão da maior fatia do orçamento destinado à saúde, conforme demonstrado na Fig. 2, que mostra os 5 favorecidos que mais aplicam recursos na saúde (UNIÃO, 2021), haja visto que, os atendimentos primários e acompanhamentos são realizados principalmente pelos municípios que tem maior contato com os cidadãos.

Figura 2 – Gráfico aplicação recursos da saúde 2021



Fonte: Adaptado de União (2021)

Em Curitiba, a rede de atenção à saúde conta com uma estrutura que compreende a

atenção básica, especializada, urgências, emergências e a realização de exames para atender pacientes do município e da região metropolitana, sendo reconhecida como referência para outros sistemas municipais de atendimento à saúde. A atenção básica conta com 111 unidades básicas de saúde e são a referência dos pacientes e porta de entrada para o SUS. Na Atenção Especializada, os atendimentos são realizados por aproximadamente 75 locais, como hospitais, ambulatórios e clínicas. A rede de Urgência e Emergência é composta por 9 UPAs que funcionam em tempo integral (CURITIBA, 2021a).

Para facilitar o gerenciamento das unidades de saúde, Curitiba tem a sua área geográfica dividida em distritos sanitários, esses agrupam a população com características epidemiológicas e sociais semelhantes (SANTOS, 2018). O município é dividido em 10 distritos sanitários (CURITIBA, 2021a), conforme mostra a Fig. 3.

Figura 3 – Distritos Sanitários de Curitiba

Fonte: Curitiba (2021a)

4.1 Base de dados do E-saúde

O sistema Informatizado E-saúde é utilizado para o registro dos atendimentos prestados pela Secretaria Municipal de Saúde de Curitiba em sua rede de atenção composta por Unidades Básicas de Saúde, Unidades de Pronto Atendimento, Centros de Especialidades Médicas e Odontológicas. A base de dados do E-saúde é disponibilizada no portal de dados abertos logo após o encerramento do mês e cada arquivo possui os dados acumulados dos últimos três meses. Para este trabalho foi selecionado o arquivo do perfil de atendimento médico nas unidades municipais de saúde do primeiro trimestre de 2020. Também está disponível no portal de dados abertos o dicionário de dados, que contém a descrição técnica e a tipagem de cada atributo da base (CURITIBA, 2021b).

Os arquivos são disponibilizados em formato separado por ponto e vírgula (.csv) e podem ser visualizados em programas de edição de texto ou de edição de planilhas eletrônicas. O arquivo do primeiro trimestre tem 898.849 linhas e tamanho de 360 MB, com 42 atributos, contendo informações de atendimentos, dados dos pacientes, de internação e da farmácia Curitibana, conforme apresentado na tabela 1.

Dentre os 42 atributos da base em primeira análise, os mais importantes são os que contém informações dos atendimentos e dos pacientes. Em relação ao atendimento, os atributos *tipo de unidade, descrição do procedimento, descrição do CBO, descrição do CID, solicitação de exames* e *desencadeou internamento*, que em conjunto informam características e padrões nos atendimentos. Já quanto aos pacientes os atributos *sexo, bairro* e *idade* contém aspectos importantes do perfil dos pacientes que são atendidos em Curitiba.

4.2 Aplicação do KDD na Base de dados do E-saúde

Esta seção trata da aplicação do KDD sobre a base de dados do E-saúde do perfil de atendimento médico. A seção é dividida em três partes que estão descritas a seguir, sendo elas preparação dos dados, transformação e mineração de dados e interpretação de resultados.

4.2.1 Preparação dos dados ou limpeza dos dados

Na exploração inicial dos dados do E-saúde, utilizou-se a linguagem $python$ para realizar análises preliminares no sentido de avaliar a base de dados e verificar a quantidade de registros nulos e para confirmar se o tipo de dados existentes no campo é o mesmo do informado no dicionário de dados.

Morfologicamente, dos 42 atributos da tabela, 30 são do tipo texto, 3 são tipo data e 9 tem tipo numérico e são aderentes as informações de tipo que constam no dicionário de dados. No mapeamentos dos valores nulos, conforme Fig. 4, identificou-se 17 atributos com valores nulos que são apresentados na tabela 2 com a respectiva quantidade de instâncias com valores nulos.

Tabela 1 – Atributos da base de atendimento médico

Nº	Grupo	Nome do Atributo
1	atendimento	Data do Atendimento
2	paciente	Data de Nascimento
3	paciente	Sexo
4	atendimento	Código do Tipo de Unidade
5	atendimento	Tipo de Unidade
6	atendimento	Código da Unidade
7	atendimento	Descrição da Unidade
8	atendimento	Código do Procedimento
9	atendimento	Descrição do Procedimento
10	atendimento	Código do CBO
11	atendimento	Descrição do CBO
12	atendimento	Código do CID
13	atendimento	Descrição do CID
14	atendimento	Solicitação de Exames
15	farmácia	Qtde Prescrita Farmácia Curitibana
16	farmácia	Qtde Dispensada Farmácia Curitibana
17	farmácia	Qtde de Medicamento Não Padronizado
18	atendimento	Encaminhamento para Atendimento Especialista
19	atendimento	Área de Atuação
20	internamento	Desencadeou Internamento
21	internamento	Data do Internamento
22	internamento	Estabelecimento Solicitante
23	internamento	Estabelecimento Destino
24	internamento	CID do Internamento
25	paciente	Tratamento no Domicílio
26	paciente	Abastecimento
27	paciente	Energia Elétrica
28	paciente	Tipo de Habitação
29	paciente	Destino Lixo
30	paciente	Fezes/Urina
31	paciente	Cômodos
32	paciente	Em Caso de Doença
33	paciente	Grupo Comunitário
34	paciente	Meio de Comunicação
35	paciente	Meio de Transporte
36	paciente	Município
37	paciente	Bairro
38	paciente	Nacionalidade
39	paciente	cod usuário
40	paciente	origem usuário
41	paciente	residente
42	atendimento	cod profissional

[1]Fonte: Autoria própria

Figura 4 – Busca dados nulos

```
# verifica a quantidade de dados nulos para cada atributo
nulos = saude.isnull().sum()
nulos
```

Fonte: Autoria própria

No tratamento dos atributos *descrição do CID* e *código do CID* as 165 linhas com valores nulos foram removidas, conforme Fig. 5, por se tratar de poucas instâncias em relação ao total de linhas da base. Após a exclusão o arquivo ficou com o total de 898.684 linhas.

Para o atributo *Tratamento no Domicílio*, os valores nulos foram substituídos pela moda, conforme Fig. 6, que é o valor que mais aparece na coluna, foi adotada devido à coluna apresentar tipo de dado nominal. Essa técnica é aplicada para evitar distorções ou

Tabela 2 – Quantidade de atributos nulos na base de perfil de atendimento médico

Descrição atributo	valores nulos
Data do Atendimento	0
Data de Nascimento	0
Sexo	0
Código do Tipo de Unidade	0
Tipo de Unidade	0
Código da Unidade	0
Descrição da Unidade	0
Código do Procedimento	0
Descrição do Procedimento	0
Código do CBO	0
Descrição do CBO	0
Código do CID	165
Descrição do CID	165
Solicitação de Exames	0
Qtde Prescrita Farmácia Curitibana	0
Qtde Dispensada Farmácia Curitibana	0
Qtde de Medicamento Não Padronizado	0
Encaminhamento para Atendimento Especialista	0
Área de Atuação	813441
Desencadeou Internamento	0
Data do Internamento	892369
Estabelecimento Solicitante	892254
Estabelecimento Destino	892254
CID do Internamento	892254
Tratamento no Domicílio	137291
Abastecimento	137231
Energia Elétrica	0
Tipo de Habitação	137219
Destino Lixo	137224
Fezes/Urina	137225
Cômodos	27081
Em Caso de Doença	137266
Grupo Comunitário	137331
Meio de Comunicação	137298
Meio de Transporte	137269
Município	0
Bairro	0
Nacionalidade	0
cod usuário	0
origem usuário	0
residente	0
cod profissional	0

[2]Fonte: Autoria própria

Figura 5 – Remove linhas com Código do CID nulos

```
# exclui linhas com código do CID nulo
saudea = saude.dropna(subset=['Código do CID'], how='all')
```

Fonte: Autoria própria

inconsistências na aplicações dos algoritmos ao utilizá-la para substituir os nulos com os valores mais frequentes que aparecem no atributo não prejudica o resultado e torna a análise mais eficaz e aderente a realidade.

Os atributos de *Abastecimento*, *Tipo de Habitação*, *Destino Lixo*, *Fezes/Urina*, *Cômodos*, *Em Caso de Doença*, *Grupo Comunitário*, *Meio de Comunicação* e *Meio de Transporte*, que também apresentavam valores nulos e têm tipo de dado nominal, receberam o mesmo tratamento de substituição pelo valor da moda.

Figura 6 – Substituir valores nulos pela moda

```
# Substitui valor nulo pela moda
moda = saudea['Tratamento no Domicílio'].mode()[0]
saudea['Tratamento no Domicílio'].fillna(moda, inplace = True)
```

Fonte: Autoria própria

Já os atributos referente a internação dos pacientes, que são *Data do Internamento, Estabelecimento Solicitante, Estabelecimento Destino* e *CID do Internamento*, só são preenchidos quando ocorre internamento e, por isso, apresentam uma grande quantidade de linhas nulas correspondente a 99% do total da base. Assim como o atributo de Área de Atuação, que é preenchido quando ocorre encaminhamento para especialista e possui 90% de informações faltantes e, por isso, esses atributos foram descartados, conforme Fig. 7.

Figura 7 – Exclusão de colunas

```
# exclui colunas
saudea = saude.drop(['Área de Atuação',
                     'Data do Internamento',
                     'Estabelecimento Solicitante',
                     'Estabelecimento Destino',
                     'CID do Internamento'], axis = 1)
```

Fonte: Autoria própria

Após realizar a limpeza dos dados, ainda utilizando a linguagem $python$, foi aberta a conexão com o banco de dados e realizada a carga dos dados no banco chamado saúde no SGBD $PostgresSQL$, Fig. 8. Na criação da tabela no SGBD foi utilizado os mesmos nomes de atributos, mas foi necessário substituir os espaços por $underline$ (_).

O processo de carga no banco de dados demorou cerca de 15 minutos, utilizando uma máquina virtual com Ubuntu 20.04 e com 6 GB de memória RAM. O processo de importação foi interrompido uma vez devido ao separador de casas decimais nos campos de *Qtde Prescrita Farmácia Curitibana, Qtde Dispensada Farmácia Curitibana* e *Qtde de Medicamento Não Padronizado* ser formado por uma vírgula (,), sendo que o SGBD utiliza como separador o ponto (.). Para essa correção foi realizada substituição da vírgula por ponto no arquivo, conforme Fig. 9.

Após o erro na importação causando a parada do processo, a tabela foi apagada e o processo de importação reiniciado para garantir que todos os dados recebessem o mesmo tratamento.

Figura 8 – Carregar dados no SGBD

```
#popula o banco
con = pg.connect(dbname = 'saude',
                 user = 'postgres',
                 host = 'localhost',
                 password = '123')

cursor = con.cursor()

def insert_esaude(dataframe):
    lista_df = dataframe.values.tolist()
    sql = "INSERT INTO esaude (data_do_Atendimento, Data_de_Nascimento, Sexo,..
    cursor.executemany(sql,(lista_df))
    con.commit()

insert_esaude(saude1t)

cursor.close()
con.close()
```

Fonte: Autoria própria

Figura 9 – Alteração separador de casas decimais

```
#altera vírgula por ponto em todo o data frame
saude1t = saudea.replace({',':'.'},regex=True)
```

Fonte: Autoria própria

4.2.2 Transformação dos dados

Para a etapa de transformação dos dados, deve-se observar o tipo dos dados que são mais adequados ao algoritmo a ser aplicado na etapa de mineração. Os atributos: *origem do usuário* e *residente*, que tinham como dado o código utilizado no banco de dados original, foram substituídos pela descrição correspondente, Fig. 10, conforme o dicionário de dados e dessa forma possibilitando a aplicação do algoritmo Apriori para obtenção das regras de associação.

Como o município de Curitiba subdivide seu território em distritos sanitários, foi criado um novo atributo, Fig. 11, para este dado estabelecido a partir do bairro do paciente. Também foi criado o atributo de idade, Fig. 12, calculado a partir da data de nascimento e em seguida a partir da idade foi criado o atributo de faixa etária, Fig. 13. Os novos atributos foram incluídos no $Postgres$ por meio de comandos em SQL, executados diretamente no $pgAdmin4$.

4.2.3 Mineração de dados da Base do E-Saúde e interpretação dos resultados

A ferramento WEKA foi escolhida para a etapa de mineração de dados por ser intuitiva e simples, além de permitir a execução de vários algoritmos diferentes com pouco esforço de reconfiguração. Isso possibilita analisar, comparar e testar soluções de forma rápida.

No WEKA, versão 3.8.5, para importação dos dados a serem minerados, foi utilizada

Figura 10 – Alterar dados atributos origem do usuário e residente

```
# altera valor atributo
saudea.loc[saudea.origem_usuario == 1 ,
           'origem_usuario'] = 'RESIDENTE NO MUNICIPIO'

saudea.loc[saudea.origem_usuario == 2 ,
           'origem_usuario'] = 'NAO RESIDENTE NO MUNICIPIO'

saudea.loc[saudea.residente == 1 ,
           'residente'] = 'COM CADASTRO NA UBS'

saudea.loc[saudea.residente == 2 ,
           'residente'] = 'SEM CADASTRO NA UBS'
```

Fonte: Autoria própria

Figura 11 – Criação do atributo distrito sanitário

```
CREATE OR REPLACE FUNCTION func_ds()
RETURNS character as
$$
DECLARE ds char;
BEGIN
    UPDATE esaude SET ds = 'PORTAO'
        WHERE bairro = 'PORTAO' OR bairro = 'AGUA VERDE'
        OR bairro = 'CAMPO COMPRIDO' OR bairro = 'FAZENDINHA'
        OR bairro = 'GUAIRA' OR bairro = 'PAROLIN' OR bairro = 'SANTA QUITERIA'
        OR bairro = 'SEMINARIO' OR bairro = 'VILA IZABEL'
        AND origem_usuario = 'RESIDENTE NO MUNICIPIO';
    UPDATE esaude SET ds = 'BAIRRO NOVO'
        WHERE bairro = 'GANCHINHO' OR bairro = 'SITIO CERCADO'
        OR bairro = 'UMBARA' AND origem_usuario = 'RESIDENTE NO MUNICIPIO';
    RETURN ds;
END
$$
LANGUAGE plpgsql;
```

Fonte: Autoria própria

a opção de conexão com o banco de dados e realizada consulta importando os dados dos atributos selecionados para aplicação do algoritmo, conforme Fig. 14.

O Apriori trabalha com tipo de dado nominal, o que o torna adequado para a base explorada que possui a maioria dos atributos desse tipo. Então, não foram utilizados os atributos de *Data_do_Atendimento, Data_de_Nascimento, Cômodos, Qtde_Prescrita_Farmácia_Curitibana, Qtde_Dispensada_Farmácia_Curitibana, Qtde_de_Medicamento_Não_Padronizado* e *idade*, por se tratarem de dados do tipo data e numérico.

A aplicação do Apriori utilizando os 29 atributos, com os parâmetros de confiança

Figura 12 – Criação do atributo idade

```
CREATE OR REPLACE FUNCTION func_idade()
RETURNS integer as
$$
DECLARE idade char;
BEGIN
    update esaude set idade = extract(YEAR from age(cast(Data_de_Nascimento as date)));
    RETURN idade;
END
$$
LANGUAGE plpgsql;
```

Fonte: Autoria própria

Figura 13 – Criação do atributo faixa etária

```
CREATE OR REPLACE FUNCTION func_faixaetaria()
RETURNS character as
$$
DECLARE faixa_etaria char;
BEGIN
    UPDATE esaude SET faixa_etaria = 'BEBE' WHERE idade <= 2;
    UPDATE esaude SET faixa_etaria = 'CRIANÇA'WHERE idade > 2 AND idade < 12;
    UPDATE esaude SET faixa_etaria = 'ADOLESCENTE' WHERE idade > 11 AND idade < 18;
    UPDATE esaude SET faixa_etaria = 'JOVEM ADULTO' WHERE idade > 17 AND idade < 31;
    UPDATE esaude SET faixa_etaria = 'ADULTO' WHERE idade > 30 AND idade < 61;
    UPDATE esaude SET faixa_etaria = 'IDOSO' WHERE idade > 60 AND idade < 81;
    UPDATE esaude SET faixa_etaria = 'MUITO IDOSO' WHERE idade > 80;
    RETURN faixa_etaria;
END
$$
LANGUAGE plpgsql;
```

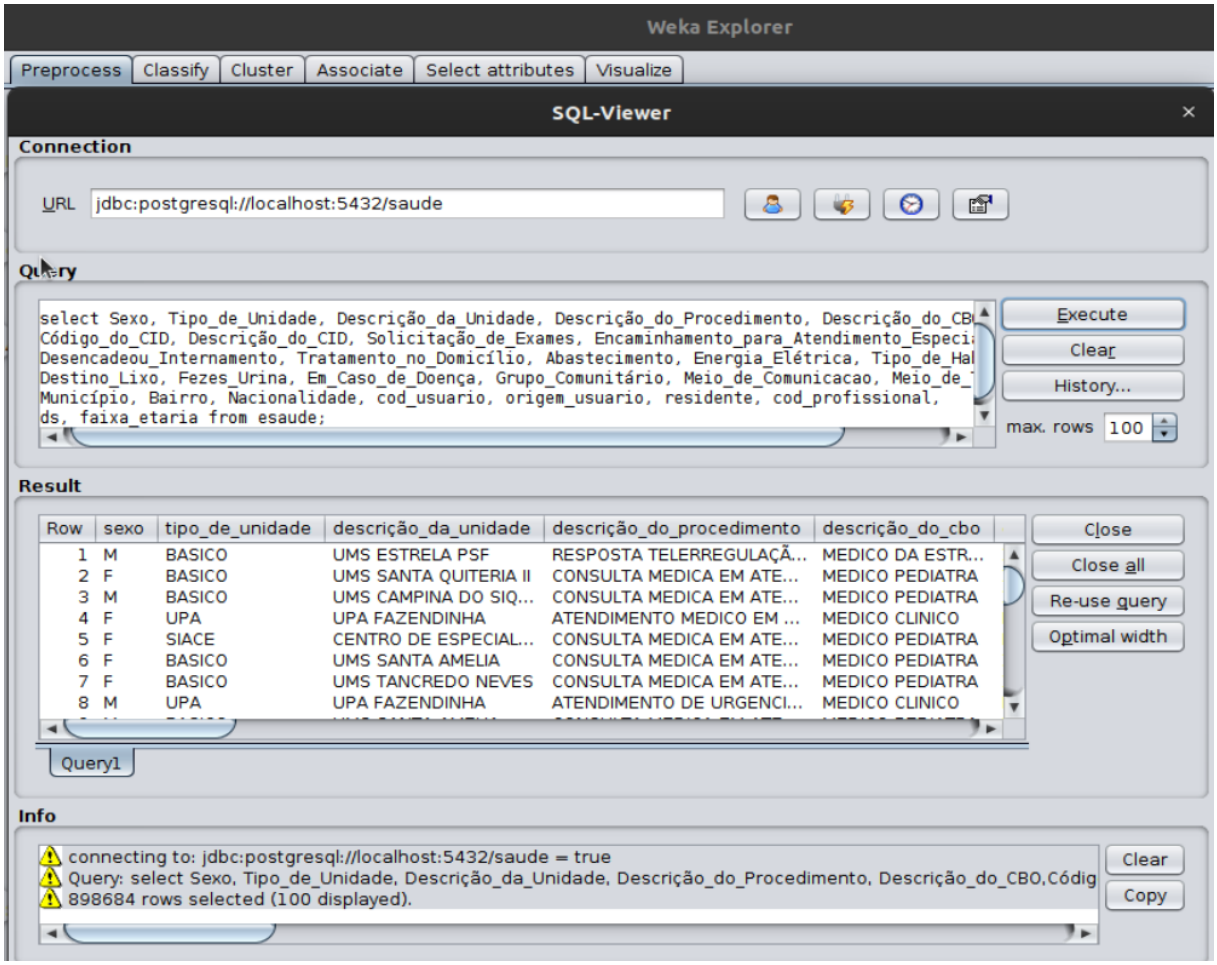
Fonte: Autoria própria

mínima de 90% e de suporte mínimo em 10% e um número de regras até 150, resultou em 150 regras após percorrer os dados por 2 ciclos (PETROSKI, 2021). Com o resultado obtido verifica-se a recorrência de regras utilizando os atributos de *nacionalidade, município* e *origem_do_usuário*. Mas, como existe uma baixa variação de valores nesses atributos, eles não geram conhecimento relevante.

Os atributos de *código_do_usuário* e *código_do_profissional* apresentam uma grande quantidade de registros exclusivos e dessa forma pouco contribuiriam para as regras de associação e, por isso, foram removidos do processo de mineração. A aplicação do Apriori utilizando os 24 atributos, com os parâmetros de confiança mínima de 90% e de suporte mínimo em 10% e um número de regras até 150, resultou em 150 regras após percorrer os dados por 3 ciclos (PETROSKI, 2021). Dada a variedade de atributos, os resultados ainda mostram regras insatisfatórias.

Continuando o processo de mineração, foi realizada um nova seleção de atributos, desta vez removendo as informações referentes ao tipo de residência e de serviços disponíveis

Figura 14 – Consulta no banco de dados através do WEKA



Fonte: Autoria própria

no domicílio do paciente que aparecem nas regras dos processamentos anteriores sem gerar conhecimento pertinente do ponto de vista de gestão. Repetindo a aplicação do Apriori, agora para os 14 atributos restantes, com os parâmetros de confiança mínima de 90% e de suporte mínimo em 10% e um número de regras até 150, resultou em 43 regras após percorrer os dados por 10 ciclos (PETROSKI, 2021). No resultado muitas regras vinculam o atributo *residente* que informa se o paciente tem ou não cadastro na unidade básica de saúde e a grande maioria dos pacientes tem o cadastro, então as regras obtidas não trazem informações significativas.

Para buscar novas regras com o Apriori, reduziu-se novamente o número de atributos. O atributo bairro apresenta 1187 valores distintos, isso se deve ao fato de haver atendimentos de pessoas de outros municípios que informam o seu bairro de residência, mas esses atendimentos representam apenas 3,59% do total, então o *bairro* foi desconsiderado. Foram retirados os atributos *código_do_CID* e *descrição_da_unidade*, este por especificar o nome da unidade de acordo com *tipo_de_unidade*, aquele por fazer par com o atributo *descrição_do_CID*.Também foram retirados os atributos *sexo, solicitação_de_exames, encaminhamento_para_atendimento_especialista* e *desencadeou_internamento*.

Então, com os 6 atributos remanescentes, referentes a dados de atendimento e do paciente, sendo eles *faixa_etária* e *descrição_do_CID*, os quais podem ser referencial para identificar algum tipo de problema. Combinado com esses as informações referente ao *tipo_de_unidade, descrição_do_procedimento* e a *descrição_do_CBO*, para identificar padrões no atendimento prestado ao paciente. E o atributo de *ds* para verificar existência de doenças regionalizadas e as necessidades da população. Para execução do Apriori foi utilizado os parâmetros de confiança mínima de 80% e de suporte mínimo em 10%, e um número de regras igual a 100, foram encontradas 34 regras após percorrer os dados por 18 ciclos (PETROSKI, 2021). Os resultados são apresentados na Fig. 15.

Figura 15 – Resultado da mineração de dados

Nº Regra	Antecendente	Consequente	Confiança	Lift
1	Proc = CONSULTA MAP 479631	==> Unid = BASICO 479631	100%	1,66
2	dCbo = ESTRAT SF 68429	==> Unid = BASICO 268429	100%	1.66
3	Proc = CONSULTA MAP dCbo ESTRAT SF 232950	==> Unid = BASICO 232950	100%	1.66
4	Proc = ATEND MED UPA 215173	==> Unid = UPA 215173	100%	2.77
5	Proc = CONSULTA MAP fEtaria = ADULTO 198502	==> Unid = BASICO 198502	100%	1.66
6	Proc = ATEND MED UPA dCbo = CLINICO 188250	==> Unid = UPA 188250	100%	2.77
7	Proc = CONSULTA MAP fEtaria = IDOSO 133635	==> Unid = BASICO 133635	100%	1.66
8	dCbo = ESTRAT SF fEtaria = ADULTO 112243	==> Unid = BASICO 112243	100%	1.66
9	Proc = ATEN URG OBS 24H AE 109332	==> Unid = UPA 109332	100%	2.77
10	Proc = CONSULTA MAP dCbo = CLINICO 108375	==> Unid = BASICO 108375	100%	1.66
11	Proc = CONSULTA MAP dCbo = ESTRAT SF fEtaria = ADULTO 103066	==> Unid = BASICO 103066	100%	1.66
12	Proc = ATEN URG OBS 24H AE dCbo = CLINICO 100396	==> Unid = UPA 100396	100%	2.77
13	Unid = UPA fEtaria = ADULTO 121672	==> dCbo = CLINICO 118448	97%	2.13
14	Unid = UPA fEtaria = JOVEM ADULTO 93170	==> dCbo = CLINICO 89981	97%	2.12
15	Unid = BASICO dCbo = CLINICO 112965	==> Proc = CONSULTA MAP 108375	96%	1.8
16	Unid = BASICO fEtaria = IDOSO 139892	==> Proc = CONSULTA MAP 133635	96%	1.79
17	Unid = BASICO fEtaria = ADULTO 216047	==> Proc = CONSULTA MAP 198502	92%	1.72
18	Proc = ATEN URG OBS 24H AE 109332	==> dCbo = CLINICO 100396	92%	2.01
19	Unid = UPA Proc = ATEN URG OBS 24H AE 109332	==> dCbo = CLINICO 100396	92%	2.01
20	Proc = ATEN URG OBS 24H AE 109332	==> Unid = UPA dCbo = CLINICO 100396	92%	2.86
21	dCbo = ESTRAT SF fEtaria = ADULTO 112243	==> Proc = CONSULTA MAP 103066	92%	1.72
22	Unid = BASICO dCbo = ESTRAT SF fEtaria = ADULTO 112243	==> Proc = CONSULTA MAP 103066	92%	1.72
23	dCbo = ESTRAT SF fEtaria = ADULTO 112243	==> Unid = BASICO Proc = CONSULTA MAP 103066	92%	1.72
24	dCbo = MEDICO GENERALISTA 99513	==> Unid = BASICO 90308	91%	1.51
25	dCid = EXAME MEDICO GERAL 103007	==> Unid = BASICO 92237	90%	1.49
26	Unid = UPA 324575	==> dCbo = CLINICO 288716	89%	1.95
27	Unid = BASICO 541861	==> Proc = CONSULTA MAP 479631	89%	1.66
28	Proc = ATEND MED UPA 215173	==> dCbo = CLINICO 188250	87%	1.92
29	Unid = UPA Proc = ATEND MED UPA 215173	==> dCbo = CLINICO 188250	87%	1.92
30	Proc = ATEND MED UPA 215173	==> Unid = UPA dCbo = CLINICO 188250	87%	2.72
31	dCbo = CLINICO fEtaria = JOVEM ADULTO 103287	==> Unid = UPA 89981	87%	2.41
32	dCbo = ESTRAT SF 268429	==> Proc = CONSULTA MAP 232950	87%	1.63
33	Unid = BASICO dCbo = ESTRAT SF 268429	==> Proc = CONSULTA MAP 232950	87%	1.63
34	dCbo = ESTRAT SF 268429	==> Unid = BASICO Proc = CONSULTA MAP 232950	87%	1.63

Fonte: Autoria própria

Para favorecer a visualização das regras na Fig. 15, foram utilizadas abreviações para os nomes dos atributos e de seus valores, cujos significados são mostrados na tabela 3.

A seguir serão analisadas as regras que apresentaram informações mais representativas.

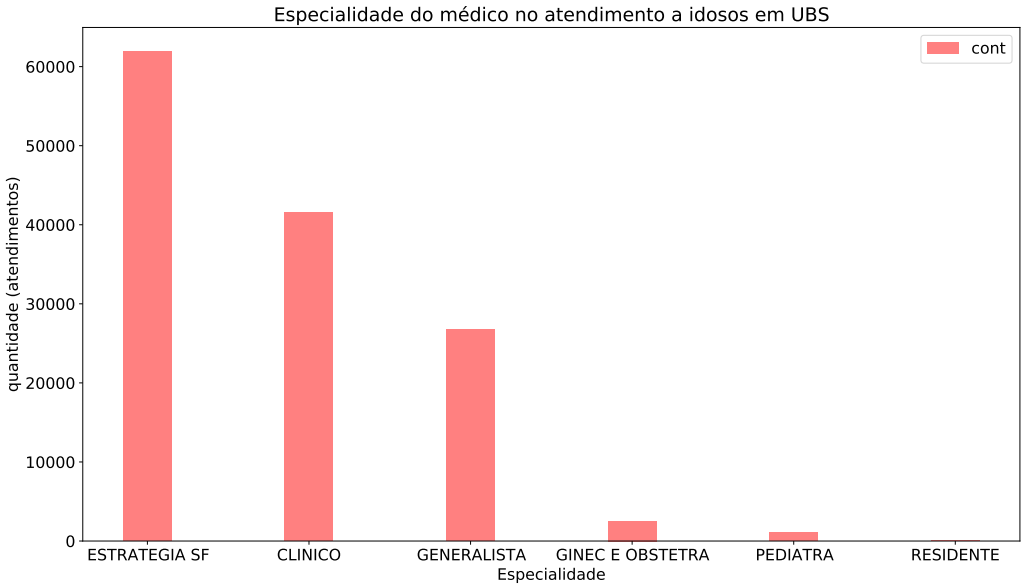
A regra 7 evidencia que 133.635 consultas médicas em atenção primária para faixa etária de idosos ocorre em unidade básica com confiança de 100%. Partindo dessa regra podemos verificar qual a especialidade do médico que efetuou os atendimentos, conforme Fig. 16.

Tabela 3 – Abreviações de nome dos atributos e de seus valores

Abreviação	Descrição original
ATEN URG OBS 24H AE	Atendimento de urgência c/ observação até 24 horas em atenção especializada
ATEND MED UPA	Atendimento médico em Unidade de Pronto Atendimento
CLINICO	Médico clínico
CONSULTA MAP	Consulta médica em atenção primária
dCBO	descrição do cbo
dCid	descrição do CID
ESTRAT SF	Médico da estratégia de saúde da família
fEtaria	faixa etária
Proc	descrição do procedimento
Unid	tipo de unidade

[3]Fonte: Autoria própria

Figura 16 – Gráfico especialidade do médico no atendimento a idosos nas unidades básicas de saúde



Fonte: Autoria própria

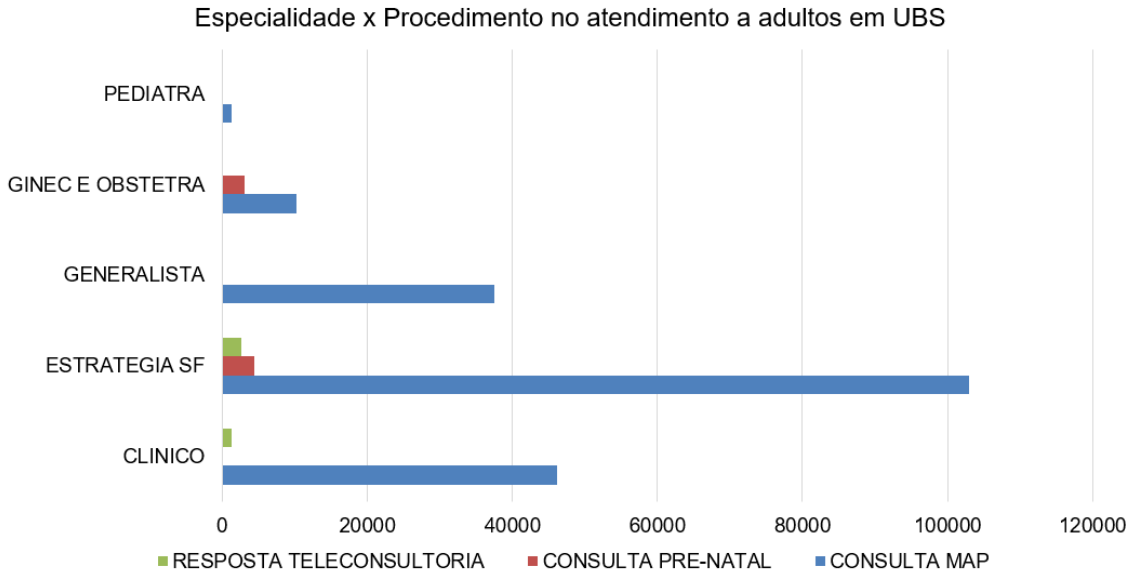
Diante da combinação da regra 7 com a *descrição_do_CBO* o gestor poderia disponibilizar nas unidades de saúde médicos com especialidade geriátrica, trazendo mais qualidade ao atendimento da população idosa. No gráfico ainda fica evidente que ocorrem atendimentos a idosos por médico com especialidade em pediatria.

Na regra 8 constata-se que 112.243 atendimentos prestados por médico da estratégia da família para a faixa etária adulto ocorrem em unidades básicas com confiança de 100%. Em complemento a regra 11 demonstra que 103066 atendimentos são para consulta médica de atenção primária, também com confiança de 100%. Enquanto as regras 21, 22 e 23 também relacionam atendimentos prestados por médico da estratégia da família para a faixa etária adulto que ocorrem em unidades básicas no procedimento de consulta médica de atenção primária, mas essas com confiança de 92%.

Com base nas regras 8, 11, 21, 22 e 23 obtemos o gráfico demonstrado na Fig.

17 observando *descrição_do_CBO* versus a *descrição_do_procedimento* evidenciando que há atendimentos a adultos por outras especialidades e com outros procedimentos, para melhor visualização no gráfico foram considerados especialidades com mais de 1000 atendimentos. Um ponto que chama atenção é que há médicos com especialidade de estratégia da família atendendo consulta pré-natal. Com base nessa informação, o gestor pode acompanhar se os médicos estão atuando na sua especialidade e se há necessidade de realocação de pessoal ou das atividades executadas.

Figura 17 – Gráfico especialidade x procedimento no atendimento a adultos nas unidades básicas de saúde



Fonte: Autoria própria

A regra 9 demonstra que 109.332 atendimentos de urgência com observação de até 24 horas em atenção especializada são prestados nas UPAs, com confiança de 100%. Na regra 12 temos que 100396 dos atendimentos da regra 9 foram realizados por médico clínico com confiança de 100%, resultado semelhante é encontrado nas regras 18, 19 e 20, porém estas com confiança de 92%, pois mesclam as associações encontradas nas regras 9 e 12. Com base nessas informações o gestor pode avaliar a necessidade e possibilidade de realocação de espaços hospitalares reservados para este fim, já que há no município 9 UPAs (CURITIBA, 2021a), mas o território é dividido em 10 distritos sanitários, deixando descoberto o distrito sanitário da Matriz, o que implica em maior deslocamento do paciente para receber atendimento.

Com base nas regras analisadas, verificou-se que a aplicação do KDD, com o algoritmo Apriori na etapa de mineração de dados, pode ser uma ferramenta interessante de suporte à gestão, em particular à gestão em saúde municipal. Na base de dados avaliada, do E-saúde, foram localizadas diversas relações entre os atributos selecionados evidenciando a especialidade do médico que prestou o atendimento relacionando com o procedimento realizado, a faixa etária dos pacientes e o tipo de unidade de saúde.

As regras obtidas revelam que as unidades básicas de saúde atendem a um grande número de idosos mesmo sem ter um médico especialista em geriatria, sendo um ponto de avaliação importante para melhorar a qualidade dos atendimentos e as políticas públicas de atendimento a essa população mais vulnerável.

Além disso, foi localizado um possível desvio entre as especialidades do médico e o atendimento realizado. A partir das regras com a combinação de outros atributos foram criados gráficos que evidenciam que há atendimento a idosos por médico pediatra e consultas pré-natal que são realizadas por médicos que não tem especialidade de obstetrícia, explicitando a necessidade da gestão rever a alocação de pessoal bem como a distribuição de atividades para evitar desvios de especialidade.

Nos experimentos conduzidos, o KDD foi aplicado apenas nos dados relativos ao primeiro trimestre de 2020. Embora, com os atributos usados, tenham sido identificadas informações úteis e relevantes para a gestão, estima-se que outras tantas informações ainda sejam passíveis de descobertas via novas formas de exploração da base. Recomenda-se, ainda, que as regras obtidas sejam analisadas pelos gestores de saúde municipal e especialistas da área quanto a sua qualidade, aplicabilidade e correspondência com os processos reais, a fim de que de fato possam subsidiar tomadas de decisões afirmativas.

5 CONCLUSÃO

Os milhares de dados que são gerados atualmente precisam de ferramentas adequadas para armazenamento, processamento, manipulação e para extração de informação e conhecimento para melhor percepção e tomada de decisão. Na área de saúde o KDD pode revelar nos dados padrões importantes para auxiliar profissionais da saúde e os gestores públicos.

O KDD é um método poderoso para descoberta de novas informações em grandes conjuntos de dados, permite que o usuário se concentre nas informações relevantes em meio a uma tempestade de dados. O KDD passa por cinco etapas: seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados. Neste trabalho foi aplicado o KDD na base de dados de perfil de atendimento médico do sistema E-saúde do município de Curitiba.

Na etapa de pré-processamento e limpeza dos dados através de $pipeline$ do $python$ foi analisado o tipo de dado dos atributos e verificada a existência de atributos com valores nulos. Esses foram tratados através de exclusão de linhas e substituição de nulos pela valor da moda do atributo, uma vez que essa técnica encontra os valores mais frequentes e dessa forma melhora a qualidade da base para possibilitar a aplicação do algoritmo na etapa de mineração de dados. Ainda foram excluídas cinco colunas com dados sobre internações, pois apresentavam mais de 90% de linhas nulas.

Na etapa de transformação observando a aplicação do algoritmo Apriori foram tratados dois atributos que possuíam o código do banco de dados original substituindo esse pela sua descrição. Além disso, foram criados novos atributos a partir dos dados já existentes na base com as informações de idade, faixa etária e distrito sanitário.

E, na mineração de dados, foi aplicado através do WEKA o algoritmo Apriori para mapeamento de padrões por regras de associação. Foram selecionados os atributos com tipo de dados nominais e realizados quatro processamentos com conjuntos de atributos diferentes. O último processamento com o conjunto de seis atributos, sendo eles procedimentos do atendimento, descrição do CID, tipo de unidade, especialidade do médico que prestou o atendimento, o distrito sanitário e a faixa etária do paciente, resultou em 34 regras com resultados pertinentes, após percorrer os dados por 18 ciclos utilizando os parâmetros de 80% para confiança mínima e 10% para suporte mínimo.

Esses seis atributos são relevantes para os gestores avaliarem a distribuição de recursos, além de proporcionar informações sobre o perfil do público ao qual unidade presta seus serviços, pode com base nesses dados fazer escolhas direcionadas as demandas buscando supri-las.

Finalmente, na interpretação dos resultados foram avaliadas as regras mais significativas do ponto de vista de gestão. Nessa etapa fica evidente que o Apriori encontra regras no conjunto de dados explorado obtendo conhecimento útil, que deve passar por análise humana para avaliar a qualidade e aplicabilidade das informações. E ainda as regras podem ser o ponto de partida

para outras análises, por exemplo a agregação de outros atributos à regra aprofundando a obtenção de conhecimento e revelando novas informações que não estavam evidentes na regra original bem como possibilita analisar à exceção das regras que revelam informações significativas para a gestão.

Este trabalho contribui para a gestão municipal de saúde com a extração de informações úteis em relação aos atendimentos e perfil dos pacientes através da aplicação do KDD com o algoritmo Apriori. A esfera municipal é responsável pela execução da maior parcela do orçamento destinado a saúde no Brasil, diante desse fato é imprescindível uma boa gestão da rede de atenção a saúde para eficiência e eficácia na destinação dos recursos e um atendimento de qualidade aos cidadãos.

5.1 Limitações

Uma limitação encontrada no desenvolvimento deste trabalho foi em relação ao processamento computacional, devido ao volume de dados. E, por isso, a inclusão dos dados dos demais trimestres de 2020 não foi possível devido às configurações da máquina disponível. Outra limitação é em relação a disponibilidade das informações publicadas no portal de dados abertos, mas os gestores e agentes públicos da saúde municipal tendo acesso à totalidade do banco de dados do E-saúde podem dispor de uma gama maior de atributos que resultem em combinações novas com melhores resultados.

5.2 Trabalhos futuros

Trabalhos futuros podem explorar a base de dados do E-saúde buscando novas relações e podem utilizar novos atributos para obtenção de novas regras obtendo outras informações úteis. Seria interessante ter atributos com informações das queixas dos pacientes e nome do medicamento prescrito, visto que já existe a informação da quantidade de medicamento prescrita e dispensada pela farmácia Curitibana, esses atributos possibilitariam outros tipos de análises.

Referências

ABDRABO, M. et al. Enhancing big data value using knowledge discovery techniques. **IJ Information Technology and Computer Science**, v. 8, n. 1-12, p. 1–4, 2016. Citado na página 15.

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: **Proceedings of the 1993 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 1993. p. 207–216. Citado na página 19.

AGRAWAL, R.; SRIKANT, R. et al. Fast algorithms for mining association rules. In: CITESEER. **Proc. 20th int. conf. very large data bases, VLDB**. [S.l.], 1994. v. 1215, p. 487–499. Citado na página 19.

ARAÚJO, T. S.; OLIVEIRA, T.; SILVA, E. Sistemas inteligentes de apoio à tomada de decisão na gestão pública municipal: uma abordagem conceitual. In: **CONFERÊNCIA SUL-AMERICANA EM CIÊNCIA E TECNOLOGIA APLICADA AO GOVERNO ELETRÔNICO. Anais... Palmas: CONEGOV**. [S.l.: s.n.], 2007. Citado na página 11.

BAYARDO, R. J.; AGRAWAL, R.; GUNOPULOS, D. Constraint-based rule mining in large, dense databases. **Data mining and knowledge discovery**, Springer, v. 4, n. 2, p. 217–240, 2000. Citado na página 19.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1–29, 2009. Citado 4 vezes nas páginas 16, 17, 18 e 19.

CARDOSO, S. J.; SANTANA, S. d. S. **Reconhecimento de caracteres manuscritos off-line utilizando Support Vector Machine (SVM)**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2019. Citado na página 18.

CERQUIDES, J.; MANTARAS, R. Lopez de. Fuzzy metaqueries for guiding the discovery process in kdd. In: **Proceedings of 6th International Fuzzy Systems Conference**. [S.l.: s.n.], 1997. v. 3, p. 1555–1559 vol.3. Citado na página 15.

COLLAZOS, K.; BARRETO, J. M.; PELLEGRINI, G. F. Análise do prontuário médico para a utilização com kdd. In: **CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE - CBIS**. [S.l.: s.n.], 2000. Citado na página 20.

CURITIBA, P. de. **DA ATENÇÃO BÁSICA AO ATENDIMENTO ESPECIALIZADO**. 2021. Disponível em: <https://www.curitiba.pr.gov.br/noticiasespeciais/como-funciona-o-sistema-de-saude/8>. Acesso em: 12 de out de 2021. Citado 2 vezes nas páginas 22 e 33.

CURITIBA, P. de. **Dados Abertos Prefeitura Municipal de Curitiba**. 2021. Disponível em: <https://www.curitiba.pr.gov.br/conteudo/sobre/1497>. Acesso em: 15 de ago de 2021. Citado 2 vezes nas páginas 11 e 23.

ESCOBAR, L. et al. Descoberta de padrões para identificação de casos de alto custo em operadoras de planos de saúde. **Revista Stricto Sensu**, v. 4, p. 01–21, 06 2019. Citado 4 vezes nas páginas 11, 12, 13 e 21.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996. Citado 6 vezes nas páginas 12, 15, 16, 17, 18 e 20.

FEDERAL, G. **Campanha Nacional de Vacinação contra Covid-19**. 2021. Disponível em: <https://dados.gov.br/dataset/covid-19-vacinacao>. Acesso em: 05 de nov de 2021. Citado na página 21.

FEDERAL, G. **Portal Brasileiro de Dados Abertos**. 2021. Disponível em: <https://www.gov.br/governodigital/pt-br/dados-abertos/portal-brasileiro-de-dados-abertos>. Acesso em: 04 de nov de 2021. Citado na página 11.

FEDERAL, G. **Registro de Ocupação Hospitalar COVID-19**. 2021. Disponível em: <https://dados.gov.br/dataset/registro-de-ocupacao-hospitalar>. Acesso em: 05 de nov de 2021. Citado na página 21.

FEUSER, R. J. Mineração de dados com regras de associação aplicada em dados de unidade de saúde de pronto atendimento. Universidade Tecnológica Federal do Paraná, 2017. Citado 4 vezes nas páginas 12, 13, 19 e 21.

GRADY, N. W. Kdd meets big data. In: **2016 IEEE International Conference on Big Data (Big Data)**. [S.l.: s.n.], 2016. p. 1603–1608. Citado na página 15.

GREGORY, G.; PRETTO, F. MineraÇÃo de dados para descoberta de conhecimento em dados de promoÇÃo À saÚde. **Revista Destaques Acadêmicos**, v. 8, n. 4, 2016. Citado 3 vezes nas páginas 12, 13 e 21.

HIROTA, K.; PEDRYCZ, W. Fuzzy computing for data mining. **Proceedings of the IEEE**, v. 87, n. 9, p. 1575–1600, 1999. Citado 2 vezes nas páginas 11 e 15.

MACIEL, T. et al. Mineração de dados em triagem de risco de saúde. **Revista Brasileira de Computação Aplicada**, v. 7, n. 2, p. 26–40, maio 2015. Disponível em: <http://seer.upf.br/index.php/rbca/article/view/4651>. Citado 3 vezes nas páginas 12, 13 e 21.

MOHAMED, E. B.; LTIFI, H.; AYED, M. B. Using visualization techniques in knowledge discovery process for decision making. In: IEEE. **13th International Conference on Hybrid Intelligent Systems (HIS 2013)**. [S.l.], 2013. p. 93–98. Citado na página 11.

PAULA, L. C. d.; PRADO, E. F. d. Aplicação do processo de kdd em uma gestora de planos de saúde. **Revista da Iniciação científica da Libertas**, v. 2, p. 100–109, 12 2012. Citado 4 vezes nas páginas 12, 13, 20 e 21.

PETROSKI, L. de F. **Resultado da aplicação do Apriori**. 2021. Disponível em: <https://github.com/LilianPetroski/Apriori>. Acesso em: 10 de nov de 2021. Citado 3 vezes nas páginas 29, 30 e 31.

RAMZAN, M.; AHMAD, M. Evolution of data mining: An overview. In: IEEE. **2014 Conference on IT in Business, Industry and Government (CSIBIG)**. [S.l.], 2014. p. 1–4. Citado na página 12.

ROMÃO, W. et al. Extração de regras de associação em c&t: O algoritmo apriori. **XIX Encontro Nacional em Engenharia de Produção**, sn, v. 34, p. 37–39, 1999. Citado na página 19.

SANTOS, B. S. d. et al. Data mining: Uma abordagem teórica e suas aplicações. **Revista ESPACIOS| Vol. 37 (Nº 05) Ano 2016**, 2016. Citado 3 vezes nas páginas 12, 15 e 18.

SANTOS, W. H. **Estudo da base de dados abertos E-Saúde da prefeitura de Curitiba usando técnicas de mineração de dados**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2018. Citado 7 vezes nas páginas 12, 13, 14, 17, 18, 19 e 22.

SILVA, G. C. Mineração de regras de associação aplicada a dados da secretaria municipal de saúde de londrina pr. 2004. Citado na página 19.

SILVA, K. A. M. d. **Análise de perfis de doenças com base em técnicas de descoberta de conhecimento em bases de dados**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2019. Citado 7 vezes nas páginas 12, 13, 15, 16, 17, 18 e 21.

UNIÃO, C. G. da. **Órgãos que mais aplicam recursos na área de saúde e maiores favorecidos**. 2021. Disponível em: <https://www.portaltransparencia.gov.br/funcoes/10-saude?ano=2021>. Acesso em: 21 de nov de 2021. Citado na página 21.

WITTEN, I. H.; FRANK, E.; HALL, M. A. Chapter 1 - what's it all about? In: WITTEN, I. H.; FRANK, E.; HALL, M. A. (Ed.). **Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)**. Third edition. Boston: Morgan Kaufmann, 2011, (The Morgan Kaufmann Series in Data Management Systems). p. 3–38. ISBN 978-0-12-374856-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780123748560000018>. Citado 3 vezes nas páginas 11, 17 e 18.