

UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS COMO FERRAMENTA DE APOIO A GESTÃO HOSPITALAR E A TOMADA DE DECISÃO

Henrique Brancher Schmitt¹, Evandro Franzen²

Resumo: Atualmente a área da saúde gera e armazena uma enorme quantidade de dados provenientes do atendimento de pacientes. O aumento da competitividade, as dificuldades que o setor da saúde vem enfrentando nos últimos anos e a busca pelo aprimoramento no tratamento da saúde fazem destes dados fonte vital na busca por conhecimento útil na tomada de decisão. As técnicas de mineração de dados surgem como aliadas na exploração e busca de conhecimento dentro de grandes conjuntos de dados, revelando padrões e estabelecendo relações antes desconhecidos pelas instituições de saúde. O presente trabalho apresenta os conceitos e metodologias, técnicas e algoritmos de mineração de dados, bem como a aplicação do algoritmo de Redes Bayesianas para a classificação supervisionada em um Data Warehouse criado a partir da junção das bases de dados de duas instituições de saúde. As ferramentas Weka e Netica foram utilizadas com o objetivo de extrair conhecimento útil para auxiliar no processo de tomada de decisão.

Palavras-chave: Mineração de Dados. Descoberta de Conhecimento. Redes Bayesianas. Dados médicos.

1 INTRODUÇÃO

A tecnologia da informação aplicada ao trabalho permitiu que, através de suas inovações, sociedades inteiras vivenciassem grandes aumentos de produtividade e aprimoramento de processos (PEREZ, 2010). Segundo Porter e Millar (1985), a tecnologia da informação é uma ferramenta poderosa na transformação das mais diversas empresas, possibilitando a habilidade de interligar seus processos e atividades internas e externas através de sistemas integrados de gestão. Atualmente é possível observar que as tecnologias de

1 Graduado em Sistemas de Informação. Universidade do Vale do Taquari - UNIVATES.

2 Professor na Universidade do Vale do Taquari - UNIVATES. Doutor em Informática na Educação e Mestre em Ciência da Computação pela UFRGS. Graduado como Tecnólogo em Processamento de dados pela UNISINOS.

informação são utilizadas em todos os setores da economia, desde o comércio de bens e serviços até a o cuidado com a vida dentro de ambientes hospitalares.

Sistemas Integrados de Gestão (ERP - Enterprise Resource Planning) segundo a Deloitte Consulting (1998) são como um conjunto de softwares que permitem a integração e automatização da maioria dos processos da empresa com o objetivo de compartilhar dados e práticas em tempo real. As instituições de saúde sofreram nos últimos anos uma grande pressão por parte de seus profissionais para adquirir e implantar soluções tecnológicas capazes de ampliar a agilidade e segurança das informações (VALENTIM, 2010). Segundo Marin (2010), as novidades de automatização e padronização de processos trazidas ano após ano por ERPs auxiliam os profissionais de saúde no desenvolvimento de suas atividades.

ERPs hospitalares são amplamente utilizados e possuem como intuito principal melhorar o controle, tanto sobre a área gerencial quanto assistencial, o que ocasiona um grande aumento no volume de dados gerados e armazenados. Estes dados podem ser um ativo importante para a instituição. Os dados por si só não trazem nenhum benefício específico, por isso muitas instituições de saúde estão adotando outros sistemas ou métodos para tratar seus dados, tais como o BI (Business Intelligence) ou sistemas de Apoio à Decisão (SAD) e a descoberta de conhecimento através da mineração de dados.

A mineração de dados é parte de um processo conhecido como descoberta de conhecimento em bases de dados (KDD - Knowledge Discovery in Databases), que tem por objetivo aprimorar os resultados estatísticos atuais e é dividido em quatro partes principais: base de dados, preparação ou pré-processamento de dados, mineração de dados e avaliação ou validação do conhecimento (CASTRO e FERRARI, 2016). A Mineração de dados é um processo no qual homens e máquinas, cooperando entre si, exploram grandes bases de dados utilizando-se de técnicas relativamente confiáveis e validadas por sua relevância estatística, buscando por padrões e relacionamentos entre variáveis com o intuito da descoberta do conhecimento (CÔRTEZ, PORCARO e LIFSCHITZ, 2002).

O emprego das técnicas de mineração de dados na área da saúde vem ganhando grande destaque devido à enorme quantidade de dados gerados e a impossibilidade de se analisar estes dados de maneira tradicional (REDEKER, 2010; SCHEUNEMANN, 2015). Ainda segundo os autores, a mineração de dados possibilita a descoberta de informação útil a partir de grandes bases de dados.

Considerando o cenário exposto, a importância dos dados e a possibilidade de descoberta de conhecimento, este trabalho teve como objetivo combinar bases de dados de duas instituições de saúde, uma de grande e outra de médio porte. A principal questão investigada foi se a mineração de dados para descoberta de conhecimento em saúde hospitalar, aplicada sobre uma base

combinada de dois hospitais, permite extrair informações de maior relevância para servir de apoio a tomada de decisão em uma instituição de saúde.

A necessidade de aprimoramento da gestão em saúde hospitalar cresceu espontaneamente com a crise financeira pela qual o Brasil passou nos anos de 2016 e 2017, ocasião na qual alguns estados, em particular o Rio Grande do Sul, ganharam destaque por chegar a uma situação insustentável a ponto de ter de parcelar a folha de pagamento de seus servidores e atrasar os repasses de recursos para áreas vitais como saúde, segurança e educação.

Diante da situação insustentável, as instituições de saúde foram obrigadas a tomar medidas para o corte imediato de gastos e a aprimorar a qualidade da gestão hospitalar, tanto pela renegociação de contratos com convênios e municípios, quanto pela busca de informações para melhorar sua capacidade de atendimento, diminuindo custos e maximizando os recursos já disponíveis. A grande quantidade de dados coletados no atendimento de pacientes pode ser um ativo importante na busca de conhecimento para a tomada de decisão e a análise destes dados através de técnicas de mineração de dados pode solucionar as dificuldades de se estabelecer conexões entre tratamentos, diagnósticos e pacientes, agilizando o tratamento da doença, diminuindo o custo para a instituição e assim aperfeiçoando seus processos internos.

A capacidade de reconhecimento de padrões que pode ser obtida através do acesso e análise de grandes bases de dados é uma forte aliada na gestão de uma instituição e no processo de tomada de decisão (CONTRIBUTOR, 2014). A capacidade de prever situações como o tempo que o paciente tende a permanecer internado, o custo estimado que o atendimento pode gerar e o desfecho de um atendimento hospitalar podem auxiliar na redução de custos nas instituições de saúde, bem como no planejamento dos recursos técnicos e humanos.

2 REFERENCIAL TEÓRICO

Segundo o conceito adotado pela organização Mundial da Saúde (OMS) em 1994, saúde é um estado de completo bem-estar físico, mental e social, ou seja, a inexistência de transtornos físicos, psíquicos e sociais, não apenas causado pela ausência de uma enfermidade ou doença. No Brasil, nas últimas décadas, o que ocorreu na saúde foi a implementação de um sistema baseado nos hospitais nos quais a atenção é baseada principalmente em ações curativas, motivadas pela necessidade de socorro de indivíduos que já possuem uma doença instalada (MINISTÉRIO DA EDUCAÇÃO, texto digital, 2016).

Para o Ministério da Saúde do Brasil, o Hospital:

É parte integrante de uma organização médica e social, cuja função básica consiste em proporcionar à população assistência

médica integral, curativa e preventiva, sob quaisquer regimes de atendimento, inclusive o domiciliar, constituindo-se também em centro de educação, capacitação de recursos humanos e de pesquisas em saúde, bem como de encaminhamento de pacientes, cabendo-lhe supervisionar e orientar os estabelecimentos de saúde a ele vinculados tecnicamente. (MINISTÉRIO DA SAÚDE, texto digital).

Dentro do contexto da atenção à saúde, os hospitais ocupam um lugar preferencial e seu papel se distingue por reunir os requisitos de uma boa gerência que planeja, avalia, projeta, desenvolve e compete por qualidade. Os hospitais podem ser considerados o centro da gestão administrativa refinada, já que oferecem serviços de hotelaria, fomentam o espírito de humanidade e compreensão, desenvolvem pesquisas diárias; estimulam iniciativas para criação de técnicas e sistemas de atenção para debater sobre enfermidades - tudo isso com a aplicação de parâmetros exigentes de educação. (MALAGÓN-LONDOÑO, MORERA e LAVERDE, 2010).

Atualmente os hospitais recebem características consideradas como sendo de uma empresa prestadora de serviços de saúde, tendo que operar em um mundo que possui como parâmetro a qualidade do serviço prestado. Os seguintes pilares devem ser seguidos para estabelecimento dessa competência: planejamento, satisfação do trabalhador, orçamento, avaliação, retroalimentação (feedback) e ética. (MALAGÓN-LONDOÑO, MORERA e LAVERDE, 2010).

2.1 Data Warehouse

Segundo Kimball et al. (1998), um Data Warehouse não é construído para sustentar um processo funcional ou operacional, mas sim consolidar dados de formatos e fontes diferentes, facilitando assim o uso da informação. Conforme Santos e Ramos (2006), um Data Warehouse é desenhado para ser um repositório que consolida e armazena dados em um formato consistente, permitindo que ferramentas analíticas façam uma análise e exploração rigorosa dos mesmos. Ainda segundo Santos e Ramos (2006), um Data Warehouse alimentado por diversas bases de dados dos sistemas operacionais frequentemente possui grandes volumes de dados, porém as bases de dados que o suportam são mantidas de maneira independente, não afetando assim o seu desempenho.

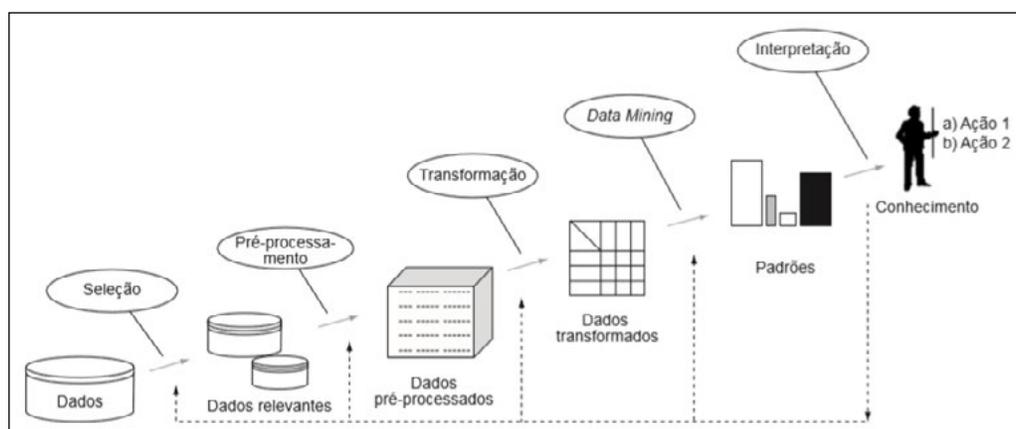
Segundo Bovo e Balancieri (2001, p. 3-7), o processo de coleta, organização e armazenamento de informações de uma ou mais bases de dados e a disponibilização adequada das mesmas para que sejam utilizadas por outro processo de análise é chamado de Data Warehousing (armazenamento de dados). Para Machado (2000), a necessidade de informações estratégicas que garantam ações rápidas na tomada de decisão está diretamente ligada a crescente utilização de Data Warehousing por empresas. Ainda segundo Machado (2000), todos os autores concordam que a tecnologia de Data

Warehousing pode ser considerada o avanço natural do universo de apoio à decisão e sua utilização é imprescindível para o aumento da competitividade em um mercado globalizado cada dia mais concorrido e variável.

2.2 Descoberta do Conhecimento

Segundo Fayyad (1996), o KDD (Knowledge Discovery in Databases ou Descoberta de Conhecimento em Bases de Dados) surgiu como alternativa para solucionar o problema de sobrecarga de dados causados pela era da informação. Ainda segundo o autor, o processo utilizado na identificação e análise de grandes bases de dados é denominado de descoberta de conhecimento em bases de dados ou KDD (Figura 1).

Figura 1 - Processo de Descoberta de Conhecimento em Etapas



Fonte Fayyad et al. (1996)

Por ser iterativo, interativo, exploratório e cognitivo, o processo de Descoberta de Conhecimento necessita ser realizado com o acompanhamento de um especialista no domínio ou análise dos dados, além disso, o processo é formado por cinco etapas, são elas:

- Na etapa de **seleção** é necessário definir quais serão os conjuntos de dados relevantes na análise da base de dados. Aqui serão avaliadas as fontes disponíveis para realizar a coleta dos dados e como estes estão estruturados.
- **Pré-processamento** é quando acontece o ajuste das informações incorretas ou inconsistentes e a eliminação dos dados não necessários com o intuito de ampliar a qualidade dos dados, essa “limpeza dos dados” trata essencialmente de dois tipos de problema, são eles:
 - **Valores ausentes** são entradas que podem não apresentar valores para alguns atributos, ou simplesmente apresentarem valores

desconhecidos, nesse caso é comum que as entradas que não possuem valores sejam preenchidas manualmente, preenchidas com alguma regra de preenchimento automático ou simplesmente removidas.

- **Valores incorretos ou com ruído** são valores que estão muito acima ou abaixo dos demais valores de um determinado atributo, ou seja, estão fora do padrão.
- **Transformação ou formatação** é a etapa na qual os dados são analisados e reorganizados de maneira que possam ser interpretados por um software de mineração de dados, para isso os dados podem ser convertidos para numéricos ou categóricos. Cabe também a esta etapa os processos de análise de frequência, binarização e discretização dos dados.
 - **Análise de frequência** objetiva agrupar ou agregar valores nos casos em que determinadas classes apareçam com pouca frequência.
 - **Binarização** é utilizada para converter valores de atributos em valores binários, por exemplo, um atributo que apresenta o custo total de uma conta pode ser convertido para 1 ou 0, representado que o custo foi alto ou baixo.
 - **Discretização** tem como objetivo a transformação de um atributo contínuo em um dado categorizado, por exemplo, um atributo que apresenta o custo total de uma conta pode ser convertido para alto ou baixo.
- **Mineração de dados**, nesta etapa utilizam-se os algoritmos de mineração de dados para transformar os dados analisados em novos conhecimentos ou novas relações entre os atributos.
- Por fim acontece a etapa de **interpretação ou avaliação**, na qual ocorre a descoberta de conhecimento através das regras indicadas na etapa de mineração de dados. Os padrões, relacionamentos e descoberta de novos fatos podem acontecer nesta etapa.

2.3 Mineração de Dados

Segundo Prather et al. (1997), mineração de dados pode ser considerado como um processo de análise de uma grande base de dados com a finalidade de encontrar padrões e relacionamentos entre dados. Carvalho (2005) complementa que com o objetivo da descoberta de conhecimento em grandes bases de dados, utiliza-se um conjunto de técnicas de estatística e inteligência artificial também conhecidas como mineração de dados. Tan, Steinbach e Kumar (2009) concordam que mineração de dados é uma metodologia capaz de proporcionar uma descoberta espontânea de conhecimento através da análise

de grandes bases de dados. Amo (2004) destaca que o processo de mineração de dados é de grande importância fazer a distinção entre o que é uma tarefa e o que é uma técnica.

Magalhães e Spínola (2014) afirmam que não existe um método universal a ser adotado; as técnicas de mineração de dados estão diretamente ligadas ao algoritmo especialista utilizado e cabe ao minerador de dados escolher qual o algoritmo que deverá ser empregado. Tan, Steinbach e Kumar (2009) sugerem uma definição breve para cada tarefa, conforme apresentado a seguir:

- **Modelagem Previsiva** é a tarefa de elaborar um padrão para a variável alvo como uma função das variáveis explicativas. A modelagem previsiva pode ser dividida em duas tarefas: classificação, quando se trabalha com variáveis discretas; e regressão, quando as variáveis são contínuas;
- **Detecção de anomalias** é a tarefa que visa distinguir argumentos nos quais as características sejam consideravelmente distintas do resto dos dados;
- **Análise de Associação** é a tarefa utilizada para revelar padrões que elucidem características fortemente associadas dentre os dados;
- **Análise de Agrupamentos (Cluster)** é a tarefa na qual se busca descobrir conjuntos de observações diretamente relacionadas, ou seja, as observações que pertencerem ao mesmo grupo serão mais semelhantes entre si do que com as que pertencerem aos demais.

São três as técnicas para desempenhar no processo de mineração de dados que são aplicadas a partir do nível de conhecimento sobre o problema que será estudado, são elas: a descoberta não supervisionada, a aprendizagem supervisionada e a modelagem matemática dos dados (CARVALHO, 2005).

Segundo Carvalho (2005), quando não existe nenhum conhecimento sobre as relações da base de dados recomenda-se a utilização da descoberta não supervisionada. Esta não possui qualquer relação predeterminada entre os dados e é utilizada para encontrar relações ocultas, não visíveis a olho nu. Ainda de acordo com o autor, após o processo, uma análise humana se faz necessária, a fim de separar as relações consideradas relevantes.

Quando já existir um conhecimento prévio das relações que estão sendo analisadas na base de dados, utiliza-se a aprendizagem supervisionada, neste caso, a mineração de dados desempenha o papel de validar ou refutar uma hipótese previamente criada (CARVALHO, 2005). Se já existir um grande conhecimento prévio da área de estudos e das relações que estão sendo analisadas, recomenda-se a utilização da modelagem matemática dos dados que permite averiguar o desempenho numérico das relações de uma afirmação com os critérios julgados interessantes (CARVALHO, 2005). A técnica representa a seleção dos algoritmos que permitam que os objetivos previamente definidos

sejam alcançados. Neste trabalho foram utilizados os algoritmos baseados em Redes Bayesianas.

Segundo Maimon, Rokach (2005), uma rede Bayesiana é constituída por dois componentes: um grafo direcionado e uma tabela de distribuição de probabilidades. As variáveis estatísticas são representadas pelos nodos dos grafos e as dependências diretas entre variáveis que são quantificadas pela probabilidade condicional de distribuição e são representadas pelos arcos. A aprendizagem em redes Bayesianas representa a indução de dois componentes distintos, isto é, a estrutura gráfica das dependências condicionais (seleção do modelo) e a distribuição condicional quantificando a estrutura de dependências (estrutura de parâmetros).

Para Witten e Frank (2005), pode-se dividir em duas etapas a aprendizagem em redes Bayesianas, sendo elas: calcular as distribuições de probabilidade e o aprendizado do grafo dirigido. A rede pode ter um processo de construção manual, conforme o conhecimento de especialistas, este processo tende a ser bastante trabalhoso caso a base de dados for grande, ou automático, podendo ser supervisionado ou não supervisionado.

Várias são as medidas para avaliar o desempenho do algoritmo de mineração de dados. As medidas adotadas neste trabalho foram a acurácia e a matriz de confusão. A matriz de confusão apresenta as classificações reais e preditas realizadas por um sistema de classificação, o desempenho do sistema é avaliado utilizando os dados da própria matriz. As classificações reais correspondem às linhas e as preditas às colunas da matriz. Caso todas as amostras tenham sido classificadas corretamente, a diagonal principal da matriz de confusão será preenchida apenas por elementos não-nulos (MARCANO-CEDEÑO et al., 2011; POLAT; GÜNEŞ, 2007).

A proximidade de uma grandeza estatística ao valor do parâmetro para o qual ela foi estimada ou seu valor verdadeiro é conhecida como acurácia (MIKHAIL e ACKERMANN, 1976 p. 44). Anzanello e Fogliatto (2011) complementam que, a acurácia define a parte da análise corretamente classificada.

3. PROCEDIMENTOS METODOLÓGICOS

Com base nos objetivos a serem atingidos, a pesquisa desenvolvida pode ser definida como exploratória. Segundo Gil (2010, p.45), a pesquisa exploratória visa possibilitar maior proximidade com o problema proposto, tornando-o mais explícito de maneira que a construção de hipóteses seja possibilitada, tendo o aprimoramento de ideias ou a descoberta de intuições como objetivo principal.

Quanto à natureza de abordagem, o método adotado foi o estudo de caso, que segundo Malhotra (2001), consiste em um processo que busca descrever e analisar certa entidade em termos qualitativos, complexos e compreensivos

e, não sistematicamente, dado que ele se desenrola em um período. O estudo de caso pode ser definido como um método qualitativo, já que as implicações a partir dos resultados obtidos não são estatísticas. Tendo como prioridade a descrição completa e o entendimento dos fatores de cada situação, o estudo de caso consiste na análise intensiva de uma ou mais situações (BOYD & STASCH, apud BARBETA, 1994).

As técnicas de mineração de dados utilizadas neste estudo com o intuito de descobrir informações implícitas em um Data Warehouse formado pela agregação de duas bases de dados de hospitais distintos são sua unidade de análise. Para realização correta deste estudo, uma amostra de 8.902 pacientes que estiveram internados em duas instituições de saúde no período de 01 de janeiro de 2017 a 30 de junho de 2018 foi utilizada. Tanto a definição do período quanto a definição dos atributos que seriam utilizados se deu em uma reunião realizada junto aos dirigentes das instituições e seus coordenadores assistenciais, com duração de aproximadamente uma hora.

A seguir iniciou-se a coleta de dados, que consiste em extrair um backup das informações de cada hospital. O sistema utilizado pelos hospitais e para o processo de backup foi o sistema Hospidata que é o ERP utilizado por ambos hospitais. O sistema Hospidata pertence ao Grupo MV e tem como objetivo integrar todas as informações hospitalares em uma plataforma única que possa facilitar o fluxo das informações entre todos os setores da empresa. Para realizar a coleta dos dados que foram utilizados no desenvolvimento deste estudo realizou-se uma busca e extração na base de dados dos hospitais com um script de consulta SQL. Após a coleta, os dados passaram por uma análise manual na qual foram limpos e normalizados para só então serem mesclados em uma base de dados única.

O processo de KDD foi iniciado pela etapa de seleção que culminou na definição de 8 atributos (Tabela 1) conforme e um total de 8.902 registros referentes às internações realizadas nas duas instituições de saúde no período de 01 de janeiro de 2017 a 30 de junho de 2018.

Tabela 1 – Definição dos atributos

Atributo	Descrição
Custo do atendimento	Indica o custo total do atendimento
Tempo de permanência	Indica o total de dias que o paciente permaneceu internado
Sexo	Indica o sexo do paciente
Idade	Indica a idade do paciente no momento do atendimento
Tipo de alta	Indica o motivo de alta do paciente
Frequência cardíaca	Indica a frequência cardíaca do paciente
Pressão	Indica a pressão arterial do paciente
Temperatura	Indica a temperatura do paciente

Fonte: do Autor (2018)

Na etapa de pré-processamento foram verificados os atributos inconsistentes nas bases de dados. Os dados incorretos foram corrigidos ou até mesmo excluídos da base com o propósito de adequá-los ao processo de Mineração de Dados. Grande parte dos erros encontrados nas bases de dados foram causados pela falta de padronização nos registros. Os atributos que mais apresentavam erros estavam diretamente ligados ao registro dos sinais vitais do paciente, como por exemplo pressão e temperatura. Também foram encontrados erros causados pela falha humana, como, por exemplo, um registro incorreto para a temperatura corporal de um determinado paciente, no qual constava uma temperatura de 262 °C, nesse caso o registro foi alterado para o valor 26.2. Na etapa de transformação realizou-se a formatação dos dados para que estes pudessem ser interpretados por algoritmos de mineração de dados.

Como exemplo de transformação é possível citar o atributo que corresponde custo do atendimento, que foi classificado em 3 faixas distintas: baixo (valores entre 0 e 299), médio (300 e 599) e alto (600 ou mais). Estes valores foram acordados em reunião com os dirigentes das instituições. Já para o atributo idade foi adotada uma regra baseada nas faixas etárias definidas pelo IBGE, que considera com criança idades entre 0 e 14, jovem entre 15 a 29, adulto entre 30 e 59 ou idoso.

Diferentemente de outros estudos que utilizam apenas um único software na etapa de mineração de dados, este fez uso de dois: o software Weka e o Netica, logo, a etapa de mineração de dados foi dividida em dois estágios: o estágio de validação e o de visualização dos atributos. O software Weka foi inicialmente utilizado para aplicação do algoritmo NaiveBayes no Data Warehouse anteriormente criado. Ao aplicar o método no Weka foi possível visualizar a rede bayesiana gerada, seus valores, sua acurácia e matriz de confusão, a qual permitiu avaliar o desempenho do algoritmo.

O software Netica, em sua versão gratuita disponível não traz informações sobre o desempenho do algoritmo, porém, apresenta uma rede bayesiana com qualidade superior a apresentada pelo Weka, permitindo visualizar os valores das probabilidades e a interação com os atributos.

4. RESULTADOS

Neste capítulo serão apresentadas os resultados da aplicação dos métodos de aprendizado de máquina durante a etapa de mineração de dados. Conforme citado anteriormente o Weka foi utilizado para realizar a verificação da confiabilidade do algoritmo e o Netica, para realizar a construção e treinamento da rede bayesiana.

4.1 Desempenho do algoritmo no WEKA

Na Tabela 2 são apresentados os valores de acurácia obtidos no primeiro estágio para cada atributo selecionado, além do total de instâncias corretamente classificadas.

Tabela 2 – Acurácia dos atributos

Atributo	Acurácia	Classificadas corretamente
grupo_alta	96,18%	8.879 instâncias
grupo_presao	85.55%	7.616 instâncias
grupo_temperatura	86.59%	7.708 instâncias

Fonte: Do Autor (2018)

Dos 8 atributos, apenas 3 apresentaram acurácia superior a 85% e foram utilizados para realização deste estudo, a saber, o atributo alta (grupo_alta), pressão arterial (grupo_pressao) e temperatura corporal (grupo_temperatura). A Tabela 3 representa as matrizes de confusão geradas para os atributos aprovados, bem como a quantidade de instâncias classificadas correta ou incorretamente.

Tabela 3 – Matriz de confusão

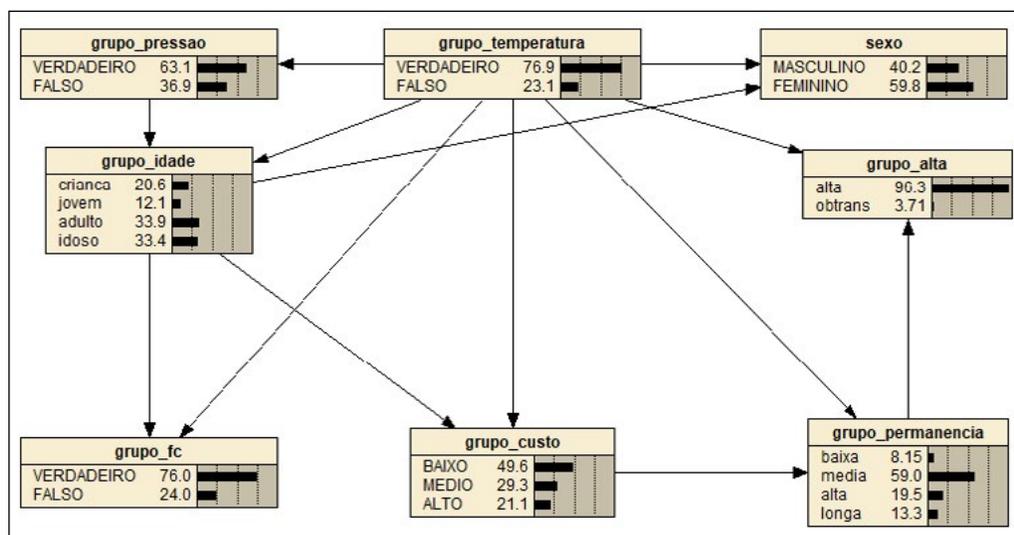
Atributo / Estado	Correto	Incorreto
grupo_alta / alta	8.559 instâncias	20 instâncias
grupo_alta / obrtrans	320 instâncias	3 instâncias
grupo_presao / verdadeiro	5.072 instâncias	541 instâncias
grupo_presao / falso	2.544 instâncias	745 instâncias
grupo_temperatura / verdadeiro	6.504 instâncias	346 instâncias
grupo_temperatura / falso	1.204 instâncias	848 instâncias

Fonte: Do Autor (2018)

4.2 Resultados obtidos com o software Netica

Partindo da suposição que as alterações na temperatura corporal do paciente possam ser um indicativo de uma infecção, resultando no uso de antibióticos e em um maior tempo de internação, realizou-se o primeiro experimento com o objetivo de verificar a ligação entre a temperatura do paciente, o custo da conta e o tempo de internação (Figura 2)

Figura 2 – Rede Bayesiana para o atributo temperatura



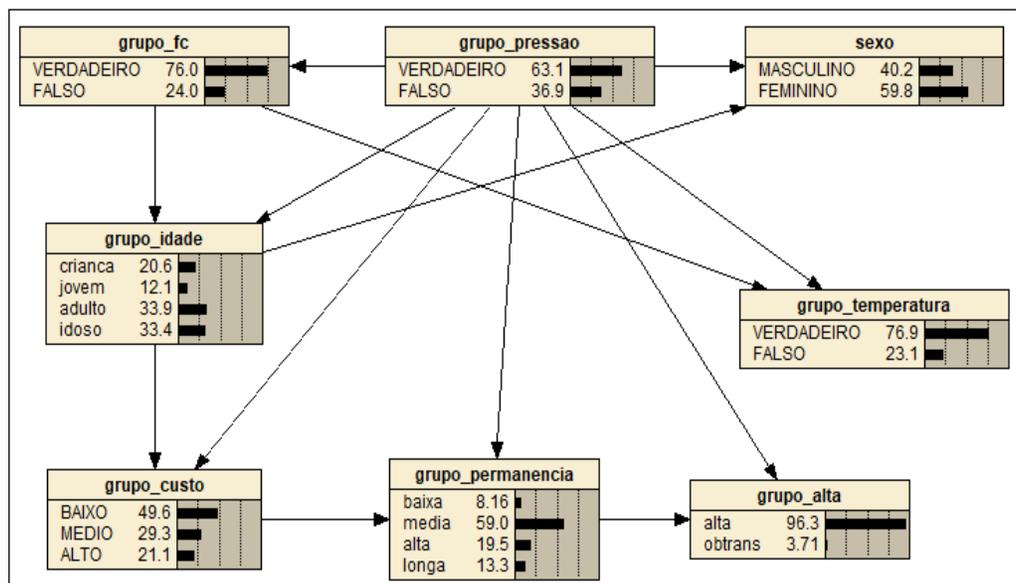
Fonte: Do Autor (2018)

Não foram identificados sinais expressivos no aumento do custo ou do tempo de atendimento dos pacientes apenas evidenciando o nodo temperatura como VERDADEIRO, porém notou-se um aumento de 10,2% de chances de o nodo pressão (grupo_pressao) ser VERDADEIRO e 11,8% de o nodo frequência

cardíaca (grupo_fc) também ser VERDADEIRO, indicando uma clara ligação entre temperatura, pressão e frequência cardíaca.

A ligação entre os três nodos anteriormente citados torna-se ainda mais clara ao evidenciar como FALSO o valor do nodo temperatura, levando a um crescimento 33,8% nas chances de o nodo pressão e de 39,3% de o nodo frequência cardíaca também serem FALSOS. A suposição inicial do primeiro experimento só é averiguada no momento em que o nodo pressão é evidenciado como VERDADEIRO e o nodo idade (grupo_idade) é evidenciado como IDOSO, o que leva a uma redução de 17,4% nas chances de o custo do atendimento (grupo_custo) ser baixo e um aumento de 7,4% de chances de o atendimento evoluir para uma longa permanência (grupo_permanencia), elevando para 44,1% as chances de o atendimento possuir uma alta ou longa permanência. No segundo experimento optou-se pelo nodo pressão como sendo o nodo raiz da rede bayesiana (Figura 3)

Figura 3 – Rede Bayesiana para o atributo pressão

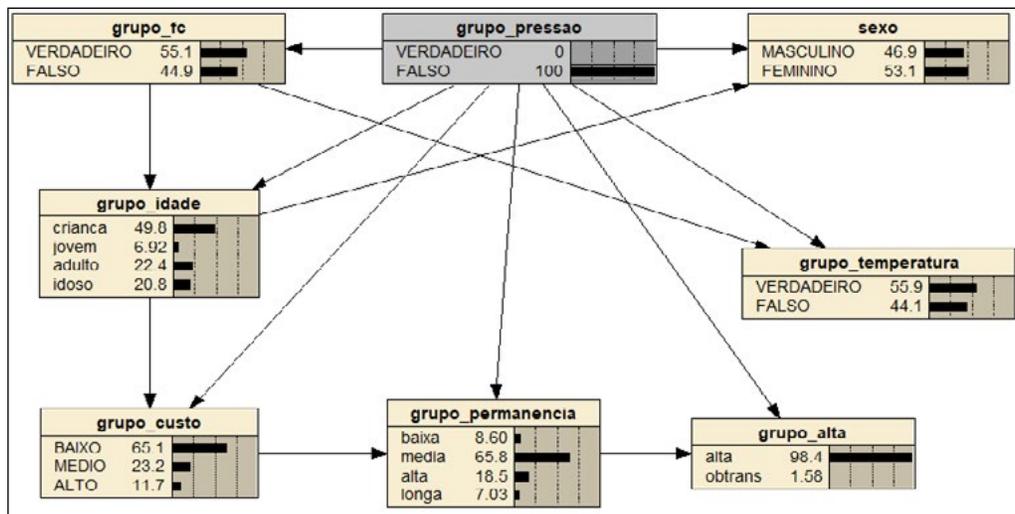


Fonte: Do Autor (2018)

Ao evidenciar como sendo VERDADEIRO o nodo pressão, observa-se um aumento de 6,7% nas chances de o paciente ser adulto e 7,4% nas chances de o paciente ser idoso, elevando para um total de 81,4% as chances de os problemas de pressão estarem relacionados a pacientes adultos ou idosos. Ao evidenciar o nodo pressão como VERDADEIRO e o nodo idade como IDOSO, percebe-se um aumento de 17,4% nas chances de o custo do atendimento não ser baixo e de 7,4% de chances de o atendimento evoluir para uma longa permanência.

Ao evidenciar o nodo pressão como FALSO, as chances de o atendimento possuir um custo baixo são de 65,1%, além disso, observa-se que existe uma chance de 49,8% dos pacientes classificados como crianças não possuírem problemas de pressão, conforme apresentado na Figura 4.

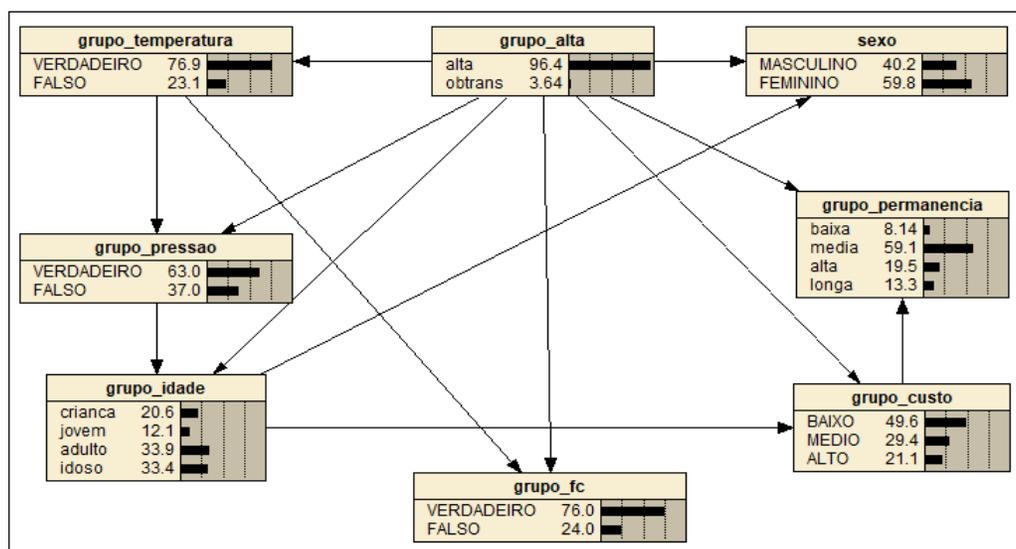
Figura 4 – Rede Bayesiana para o atributo pressão com valor FALSO



Fonte: Do Autor (2018)

Evidenciando o nodo pressão como FALSO e o nodo idade como CRIANÇA, as chances de o atendimento possuir um custo baixo aumentam para 86,1%, já as chances de o atendimento possuir baixa ou média permanência somam 87,3%. O terceiro e último experimento foi realizado utilizando o nodo alta como nodo raiz, a Figura 5 apresenta a rede bayesiana gerada a partir da aplicação do algoritmo de classificação.

Figura 5 - Rede Bayesiana para o atributo alta



Fonte: Do Autor (2018)

Ao evidenciar o nodo alta (grupo_alta) como ALTA não foram observadas variações significativas nos demais nodos, porém, ao evidenciar o mesmo como ÓBITO OU TRANSFERÊNCIA, percebe-se um aumento de 30,5% nas chances de o paciente ser idoso, elevando o valor de 33,4% para 63,9%, já as chances do atendimento possuir um custo alto são de 45,6% e as chances de o atendimento ser uma longa permanência são de 32,7%.

Caso o nodo alta seja evidenciado como ÓBITO OU TRANSFERÊNCIA e o nodo idade seja evidenciado como criança, percebe-se uma chance de 78,6% de o custo do atendimento ser baixo e um total de 78,3% de chances de o tempo de permanência ser baixo ou médio.

4.3 Análise dos resultados

A última etapa do processo de descoberta de conhecimento é o processo de interpretação dos resultados, nesta etapa aconteceu a validação do conhecimento extraído no processo de mineração de dados junto aos dirigentes das instituições de saúde e coordenadores das áreas assistenciais.

De acordo com os dados minerados para este estudo, os resultados mostram que, diferentemente do que se acredita, as alterações na temperatura corporal do paciente possuem maiores chances de impactar no aumento do custo do atendimento e o tempo de permanência na instituição caso o paciente seja idoso. Do ponto de vista dos gestores, essa é uma informação relevante por esclarecer uma crença que existe na instituição de saúde, embasada em informações consolidadas extraídas do ERP da instituição, além de poder

auxiliar no planejamento de cuidados ao paciente. A ligação observada entre os nodos temperatura, pressão e frequência cardíaca já era uma informação de conhecimento dos coordenadores assistenciais.

De acordo com os dados minerados para este estudo, os resultados indicam que existe uma chance de 65,1% de o custo do atendimento ser baixo nos casos em que não há alterações na pressão arterial do paciente. Tanto do ponto de vista dos gestores quanto dos coordenadores assistenciais, essa é uma informação relevante por se tratar de um problema que muitas vezes é causado pela falta de informação do paciente. Ações educativas com foco nos riscos a saúde causados pelas alterações na pressão arterial poderiam ser oferecidas com o objetivo de melhor informar o paciente.

Os demais resultados apresentados apenas confirmam as deduções dos coordenadores assistenciais, porém corroboram para definição do público alvo de futuras ações educativas, uma vez que existe uma chance de 81,4% dos problemas de pressão estarem relacionadas a pacientes adultos ou idosos.

Outra conclusão relevante é que nos casos em que haja óbito ou transferência de pacientes idosos, as chances de o atendimento possuir um custo alto (45,6%) e uma longa permanência (32,7%) aumentam significativamente. De acordo com os gestores das instituições, essa informação já é de seu conhecimento e se deve ao fato de que pacientes idosos com complicações que possam levar a óbito ou transferência, permanecem por maior tempo internados nas instituições, ocasionando assim maiores custos no atendimento.

5. CONCLUSÕES

O desenvolvimento deste estudo amplia a compreensão quanto a importância do uso das técnicas de mineração de dados para descoberta de conhecimento na área da saúde com foco na tomada de decisão. A importância e validade dos conceitos relacionados à mineração de dados foram evidenciados na revisão bibliográfica apresentada e podem ser expandidos para utilização em bases de outras instituições de saúde.

O algoritmo bayesiano demonstrou ser de grande valia no processo de descoberta de conhecimento aplicado à área da saúde, pois as relações por ele sugeridas entre os nodos da rede bayesiana auxiliaram na descoberta e verificação de padrões. A partir da análise dos dados coletados foi possível constatar em um primeiro experimento que, diferentemente do que se acreditava, o custo e o tempo de permanência dos pacientes na instituição só são impactados por alterações na temperatura corporal caso o paciente seja idoso.

O segundo experimento realizado, constata que o custo dos atendimentos possui uma chance de 65,1% de ser baixo nos casos em que o paciente não apresenta problemas de pressão arterial durante seu atendimento. Já o terceiro experimento realizado constatou que existe um aumento considerável

nas chances de o atendimento possuir um custo alto (56,1%) e uma longa permanência (38,4%) nos casos em que o paciente seja idoso e o motivo de alta seja óbito ou transferência. Constatação essa que não se repete nos casos em que o paciente seja criança, o que leva a chance de 78,3% de o atendimento possuir uma baixa ou média permanência e 78,6% de o custo do atendimento ser baixo.

Portanto conclui-se que os resultados obtidos através do processo de descoberta de conhecimento atendem o objetivo proposto por este estudo, além de ser possível concluir a eficiência e precisão do método de descoberta supervisionada e da combinação das ferramentas Weka e Netica, que trazem benefícios visíveis no processo de apuração da confiabilidade das técnicas de mineração de dados e da visualização da rede bayesiana gerada.

REFERÊNCIAS

ANZANELLO, M. J.; FOGLIATTO, F. S.; ROSSINI, K. Data mining-based method for identifying discriminant attributes in sensory profiling. **Food Quality and Preference**, v. 22, p. 139-148, 2011.

Ministério da Educação. **Saúde**. Disponível em: <<http://portal.mec.gov.br/seb/arquivos/pdf/livro092.pdf>>. Acesso em 29 mai. 2018.

CARVALHO, L. A. V. - **Datamining: a mineração de dados no marketing**, medicina, economia, engenharia e administração - Ciência Moderna - RJ, 2005.

CASTRO, L. N & FERRARI D. G. (2016). **Introdução a Mineração de Dados: Conceitos Básicos, Algoritmos e aplicações**, São Paulo, Saraiva: 2016.

CONTRIBUTOR, C. (2014). Castlight Voice: **How Big Data Will Help Save Healthcare** - Forbes.

DELOITTE Consulting. ERP's Second Wave: **Maximizing the Value of ERP-Enabled Processes**. Relatório de pesquisa publicado pela Deloitte Consulting, 1998.

ELMASRI, R. E. & Navathe, S. **Sistemas de banco de Dados**. Pearsoned do Brasil. 4a edição, 2005, 724p

FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory e SMYTH, Padhraic. **From Data Mining to Knowledge discovery**. American Association for Artificial Intelligence. 1996.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 5ª Ed. São Paulo: Atlas, 2010.

KIMBALL, R., Ross, M., Thornthwaite, W., & Mundy, J. **The Data Warehouse Lifecycle - Toolkit**. New York, EUA: John Wiley & Sons, 1998.

MACHADO, F. **Projeto de data warehouse: uma visão multidimensional**. São Paulo: Erica, 2000.

MAGALHÃES, Péricles; SPÍNOLA, Rodrigo O.. **SQL Magazine – Mineração de Dados: Tarefas e Técnicas**, Rio de Janeiro: Devmedia, 123ª Ed., 2014.

MALAGÓN-LONDOÑO, G. MORERA, R. G., LAVERDE, P. G., **Administração Hospitalar**. 3 Ed. Rio de Janeiro: Guanabara Koogan, 2010.

MARCANO-CEDENO, A; QUINTANILLA-DOMÍNGUEZ, J.; ANDINA, D. WBCD Breast Cancer Database Classification Applying Artificial Metaplasticity Neural Network. **Expert Systems with Applications**, v. 38, p. 9573–9579, 2011.

MARIN, H. d. (2010). **Sistemas de informação em saúde: considerações gerais**. Journal of Health Informatics,. Journal of Health Informatics, v. 2, n. 1.

MAIMON, Oded, ROKACH, Lior. **DM and Knowledge Discovery Handbook**. Ramat-Aviv, Israel: Springer, 2005.

MALHOTRA, N. K. **Pesquisa de Marketing: uma orientação aplicada**. 3ª Ed. Porto Alegre: Bookman, 2001.

MIKHAIL, E.; ACKERMAN, F. **Observations and Least Squares**. University Press of America, 1976. 497 p

PRATHER, J.C., LOBACH, D.F., GOODWIN, L.K., HALES, J.W., HAGE, M.L., HAMMOND, W.E. **Medical DM: knowledge discovery in a clinical data warehouse**. Division of Medical Informatics, Duke University Medical Center, Durham, North Carolina, USA. 1997.

PEREZ, G. &. (2010). Perez, G., & Fatores determinantes da adoção de sistemas de informação na área de saúde: um estudo sobre o prontuário médico eletrônico. **RAM. Revista de Administração Mackenzie (Online)**, pp. 11(1), 174-200.

POLAT, K.; GÜNEŞ, S. Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, v. 17, p. 694-701, 2007.

PORTER, Michael E. e Millar, Victor E.. How information gives you competitive advantage. **Harvard Business Review**, Julho/Agosto 1985, p.149-160.

REDEKER, G. A.. **Descoberta de Conhecimento na área de Cardiologia**. Lajeado. Dissertação de graduação do Centro Universitário UNIVATES, 2010.

SANTOS, M; RAMOS, I.. **Business Intelligence : tecnologias da informação na gestão de conhecimento**. FCA - Editora de Informática.

SCHEUNEMANN, F. **Mineração de dados para descoberta de conhecimento na área de oncologia**. Lajeado. Dissertação de graduação do Centro Universitário UNIVATES, 2015.

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

VALENTIM, M. **Gestão Mediação e uso da informação**. São Paulo : Cultura Acadêmica, 2010.

YIN, R. K. **Case study research: design and methods**. London: Sage, 1984.

WITTEN, Ian H., FRANK, Eibe. DM: **Practical Machine Learning Tools and Techniques, Second Edition**. San Francisco: Morgan Kaufmann Publishers, 2005.