UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA EM SAÚDE

LEONARDO SILVA VIANNA

FLORIANÓPOLIS
2019

Leonardo Silva Vianna

MINERAÇÃO DE DADOS PARA A PREDIÇÃO DE MORBIDADE HOSPITALAR

Dissertação submetida ao Programa de Mestrado Profissional em Informática em Saúde da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Informática em Saúde.
Orientador: Prof. Dr. Raul Sidnei Wazlawick

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Vianna, Leonardo Silva
Mineração de dados para a predição da morbidade hospitalar / Leonardo Silva Vianna ; orientador, Raul Sidnei Wazlawick, 2019.
94 p.

Dissertação (mestrado profissional) - Universidade Federal de Santa Catarina, Centro de Ciências da Saúde, Programa de Pós-Graduação em Informática em Saúde, Florianópolis, 2019.

Inclui referências.

1. Informática em Saúde. 2. Mineração de dados. 3. Previsões. 4. Morbidade. I. Wazlawick, Raul Sidnei. II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Informática em Saúde. III. Título.

Leonardo Silva Vianna

Mineração de dados para a predição de morbidade hospitalar

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Profa. Dra. Gabriela Marcellino de Melo Lanzoni
Universidade Federal de Santa Catarina

Profa. Dra. Vania Bogorny
Universidade Federal de Santa Catarina

Prof. Dr. Raul Sidnei Wazlawick
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Informática em Saúde.

Grace Teresinha Marcon Dal Sasso:54406854991
Assinado de forma digital por Grace Teresinha Marcon Dal Sasso:54406854991
Dados: 2019.09.24 12:43:00 -03'00'

_____

Profa. Dra. Grace Teresinha Marcon Dal Sasso
Coordenadora do Programa

RAUL SIDNEI WAZLAWICK:71172084904
Assinado de forma digital por RAUL SIDNEI WAZLAWICK:71172084904
Dados: 2019.09.24 10:02:43 -03'00'

_____

Prof. Dr. Raul Sidnei Wazlawick
Orientador

Florianópolis, 11 de setembro de 2019

Esse trabalho é dedicado aos meus pais e irmão, Aroldo, Maria Aparecida e Fabrício, e à minha família, Patrícia e Tomé, aqueles que sempre acompanham e apoiam minhas jornadas.

AGRADECIMENTOS

Minhas manifestações, mesmo de forma simples e breve, de agradecimento a todos que contribuíram com esse trabalho.

Ao Professor Doutor Raul Sidnei Wazlawick, pela sua orientação e seus ensinamentos, que permitiram a construção da caminhada para realização dessa pesquisa.

Aos docentes, discentes e colaboradores do Programa de Mestrado Profissional em Informática em Saúde, pelo aprendizado alcançado, pela acolhida em um ambiente novo e, principalmente, pelo conhecimento compartilhado.

À Professora Doutora Gabriela Marcellino de Melo Lanzoni e Professora Doutora Vania Bogorny, que gentilmente aceitaram participar da banca de defesa da dissertação e proveram colaborações importantes para o aprimoramento do trabalho. À Professora Doutora Sayonara de Fatima Faria Barbosa, pela discussão de ideias e incentivos.

A Deus, a quem sempre agradeço.

Todo problema é uma oportunidade disfarçada.

(John Adams)

RESUMO

A crescente demanda por serviços de saúde em hospitais produziu desafios significantes para seus gestores. Variáveis com alto grau de incerteza, como a quantidade de pacientes e a duração de seus tratamentos, agregam complicações aos processos de planejamento e dificultam o adequado cumprimento das estratégias já estabelecidas. Controlar e identificar fatores que afetam o processo de gerenciamento de unidades hospitalares depende da análise de bancos de dados de saúde. Esses, porém, possuem características complexas, dinâmicas e heterogêneas, que exigem a aplicação de métodos e ferramentas apropriadas para permitir sua adequada interpretação. Os bancos de dados dos sistemas de saúde arquivam informações valiosas, que podem ser utilizadas para aprimorar os mecanismos de gestão e melhorar a tomada de decisão. Desta forma, é importante considerar a possibilidade de prospecção de conhecimento útil a partir dos dados armazenados. O objetivo desta pesquisa é avaliar a predição de morbidade hospitalar, através da aplicação de diferentes métodos de mineração de dados nos registros de procedimentos ambulatoriais e hospitalares, obtidos dos bancos de dados de saúde pública do Brasil. O método da pesquisa consiste na execução de uma mineração de dados preditiva com a aplicação de algoritmos de aprendizagem supervisionada, para a modelagem de um problema de regressão. O maior coeficiente de correlação ρ de Pearson, individualmente obtido no intervalo de tempo de predição de três meses, através do método de mineração de dados que aplicou o algoritmo Random Forest associado com um algoritmo de seleção de atributos, no grupo de doenças do capítulo XVI do CID-10 (Algumas afecções originadas no período perinatal), foi de 0,9682. Diferentes resultados médios foram alcançados dependendo do método aplicado, do grupo de doenças analisado e do intervalo de tempo de predição proposto, os quais possibilitaram concluir que a mineração de dados nos registros ambulatoriais e hospitalares permitiu a predição da morbidade hospitalar. As predições da morbidade hospitalar obtidas podem minimizar o efeito indesejado da aleatoriedade da demanda por serviços de saúde no processo de tomada decisão. O conhecimento gerado pela mineração de dados executada nessa pesquisa pode subsidiar o adequado planejamento na gestão hospitalar, conduzindo hospitais, públicos e privados, ao equilíbrio financeiro desejado e à melhora do nível de qualidade dos serviços prestados aos pacientes.

Palavras-chave: Mineração de dados. Previsões. Morbidade.

ABSTRACT

Growing demand for hospital health services has brought significant challenges for their managers. Variables with an uncertainty high degree, such as the number of patients and the duration of their treatments, hinders the planning processes and make it difficult to properly comply with the established strategies. Controlling and identifying factors that affect the hospital unit management process depends on health database analysis. However, health information has complex, dynamic and heterogeneous characteristics, that require the appropriate application of methods and tools for its correct interpretation. Healthcare database stores valuable information useful for improvement of the management mechanisms and the decision making by healthcare professionals. Therefore, it is important to consider the possibility of prospecting useful knowledge from the stored data. The objective of this research is to evaluate the hospital morbidity prediction through different data mining methods on ambulatory and hospital procedure records obtained from public health databases in Brazil. The research method consists of performing a predictive data mining by applying supervised learning algorithm in a regression problem. The highest Pearson correlation coefficient individually obtained in the three-month prediction time interval, through the data mining method that applied Random Forest associated with an attribute selection algorithm on the disease group of the ICD-10 chapter XVI (Certain conditions originating in the perinatal period), was 0.9682. Different results were achieved depending on the method applied, the group of diseases analyzed and the proposed prediction time interval, what led to the conclusion that data mining on ambulatory and hospital records allowed the prediction of hospital morbidity. The hospital morbidity predictions obtained can minimize the undesired effect of the demand randomness for health services in the decision-making process. The knowledge generated by the data mining performed on this research can support the proper planning in hospital management, leading hospitals, public and private, to the desired financial balance and the improvement of the quality of services provided to patients.

Keywords: Data mining. Forecasting. Morbidity.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MATLAB	Matrix Laboratory
WEKA	Waikato Environment for Knowledge Analysis
KDD	Knowledge Discovery in Databases
CRISP-DM	Cross Industry Standard Process for Data Mining
kNN	k-Nearest Neighbors
SVM	Support Vector Machine
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
CID-10	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde – 10° revisão
OLAP	Online Analytical Processing
OLAM	Online Analytical Mining
SIHSUS	Sistema de Informações Hospitalares do Sistema Único de Saúde
DATASUS	Departamento de Informática do Sistema Único de Saúde
TABWIN	Tabulador de Dados para Windows
SIASUS	Sistema de Informações Ambulatoriais do Sistema Único de Saúde
PA	Produção Ambulatorial
APAC	Laudo Médico para Procedimentos de Alta. Complexidade
BPA-C	Boletim de Produção Ambulatorial Consolidado
BPA-I	Boletim de Produção Ambulatorial Individualizado
RASS-AD	Registro das Ações Ambulatoriais de Saúde da Atenção Domiciliar
RASS-PSI	Registro das Ações Ambulatoriais de Saúde da Atenção Psicossocial
RD	Reduzidas
AIH	Autorização para Internação Hospitalar
CSV	Comma Separated Values
ARFF	Attribute-Relation File Format

SUMÁRIO

1 INTRODUÇÃO

Dados relacionados a eventos de saúde são gerados através da captação de sinais e imagens do corpo humano, utilizados para a consecução de diferentes exames diagnósticos. Outros são provenientes da evolução clínica nos prontuários de pacientes, realizada por profissionais de saúde. Ainda, existem dados que são oriundos da escrituração das informações geradas em atendimentos ambulatoriais e hospitalares, públicos e privados, necessária para o registro dos procedimentos realizados e para sua disseminação obrigatória; incluídas, nesses últimos, as morbidades correlacionadas e as notificações epidemiológicas compulsórias.

A informação em saúde tem sido utilizada para atividades de monitoramento e avaliação, aplicando-se historicamente técnicas estatísticas e de análises gráficas (CHAN et al. 2017). As análises podem ajudar na detecção precoce de doenças, na predição da trajetória de enfermidades e na detecção de fraudes. Da mesma forma, permitem a identificação de eventos de baixa frequência, que normalmente não seriam detectados, mas que possuem significante impacto (MEHTA; PANDIT, 2018). A associação da informação de saúde em sistemas de suporte a decisão tem o potencial de estabelecer novos modelos de trabalho, os quais são capazes de conduzir profissionais de saúde a introspecções e prover um guia para a prática clínica (WANG et al., 2018).

Modelos preditivos podem ser utilizados como ferramenta de suporte à decisão em atividades administrativas, permitindo que gestores executem de forma mais eficaz atividades de planejamento e gerenciamento de recursos em saúde (GRAHAM et al., 2018). A aplicação de modelos preditivos, obtidos como resultado da execução de diferentes processos em mineração de dados, pode prover importante conhecimento para profissionais de saúde, que necessitam do desenvolvimento de estratégias eficientes em contextos de limitada disponibilidade de recursos. Modelos preditivos podem ser desenvolvidos utilizando bases de dados governamentais, empregando algoritmos de aprendizagem de máquina (SHIMODA et al., 2018).

O conhecimento gerado através da mineração de dados preditiva pode ser utilizado na epidemiologia, em sistemas de suporte à decisão clínica e no gerenciamento individual de cuidados de saúde; consequentemente, permitindo o aprimoramento da qualidade dos serviços de saúde, como também reduzindo custos operacionais (LINDSAY et al., 2019). Em razão da crescente demanda por serviços de saúde, unidades hospitalares necessitam de um gerenciamento adequado de seus recursos, com objetivo de garantir eficiência econômica e atendimento apropriado aos pacientes (KOPPKA et al., 2018). Custos crescentes em serviços

de saúde demandam um gerenciamento mais sistemático, que pode ser caracterizado por um esforço em medir e gerenciar o valor e a responsabilidade, bem como os resultados e os custos, principalmente em sistemas de saúde altamente fragmentados (MOSES et al., 2013). A aplicação de modelos de dados aprimora o processo de decisão estratégica, importante para o equilíbrio financeiro de um hospital (FREEMAN et al., 2018).

Usualmente obtidos através dos conceitos utilizados em Inteligência Artificial, os modelos têm o propósito de realizar previsões sobre instâncias de um dado objetivo, que pode ser elucidado através de uma função parametrizada $f(x) = y$ desenvolvida com utilização de diferentes técnicas e algoritmos (ANGERMUELLER et al., 2016; BELLINGER et al., 2017; SHI et al., 2018; SHICKEL et al., 2018).

Diversas pesquisas científicas têm aplicado processos de mineração, utilizando algoritmos de aprendizagem de máquina, para a descoberta de conhecimento em banco de dados. Pesquisadores utilizam-se de diferentes plataformas ou programas para a execução de processos de mineração de dados. A literatura científica evidencia a utilização da linguagem Python (HUNTA et al., 2018; MARTÍNEZ-GARCIA et al., 2018; ROTH, 2018) e R (GRAHAM et al., 2018; SAHNI et al., 2018); como também os aplicativos: Ariana (BAURIN et al., 2017), Bayesian Lab (KUNJIR et al., 2017), IBM SPSS (LI et al., 2018; LUO et al., 2018), Matrix Laboratory (MatLab) (KALAISELVI; SUJARANI, 2018; LI et al., 2018; VAUGHN et al., 2018), Pentaho (FREIRE et al., 2015), Rapid Miner (DAQQA et al., 2017), Waikato Environment for Knowledge Analysis (Weka) (AGRAWAL et al., 2016; NAVAZ et al., 2016; RAU et al., 2016; AFZAL et al., 2017; ALTHUNAYAN et al., 2017; AWAD et al., 2017; BIRJALI et al., 2017; CASTALDO et al., 2017; CHAN et al., 2017; CHOWDHURY et al., 2017; KUMAR; KHATRI, 2017; KUNJIR et al., 2017; QUDSI et al., 2017; ALMADANI; ALSHAMMARI, 2018; BAŞAR; AKAN, 2018; FARZI et al., 2018; MLAKAR et al., 2018; SANTOS; CARVALHO, 2018; JHA et al., 2019).

Nesses ambientes de pesquisa são utilizados distintos métodos e algoritmos para a mineração de dados, de maneira isolada ou associados entre si. Dos variados resultados de performance, utilizando as diferentes ferramentas de mineração de dados, infere-se que não existe um método melhor. Os desempenhos são dependentes do tipo de população, das variáveis analisadas e dos propósitos a serem estudados (AWAD et al., 2017).

1.1PROBLEMA

A crescente demanda por serviços de saúde em hospitais produziu desafios significantes para seus gestores. Conquanto, esses desafios envolvem altos custos associados a orçamentos e recursos limitados (SITEPU et al., 2018).

Uma inadequada gestão hospitalar resulta em desperdício de recursos importantes e produz um impacto negativo no nível do serviço de saúde ofertado aos pacientes. O gerenciamento torna-se ainda mais complicado em um ambiente de escassez orçamentária e com grande variabilidade no fluxo de pacientes. Em unidades hospitalares que possuem serviços de pronto atendimento, que se caracteriza pela aleatoriedade da demanda, o planejamento para alocação de recursos torna-se ainda mais difícil (MA; DEMEULEMEESTER, 2013).

Variáveis com alto grau de incerteza, como a quantidade de pacientes e a duração de seus tratamentos, agregam complicações aos processos de planejamento e dificultam o adequado cumprimento das estratégias já estabelecidas (FREEMAN et al., 2018). Essa situação pode resultar em alterações de curto prazo dos planos existentes, que podem afetar as expectativas das diferentes partes interessadas (gestores, fornecedores, profissionais de saúde e equipes de suporte), culminando na tomada de medidas corretivas indesejáveis (KOPPKA et al., 2018).

A aleatoriedade na demanda por serviços de saúde é indutora de ineficiência na prestação de cuidados de saúde (KUMAR; ANJOMSHOA, 2019).

Esses problemas conduzem a uma superlotação dos serviços hospitalares, produzindo consequências negativas para a saúde dos pacientes, pois aumenta o risco de mortalidade. Também, afeta negativamente a execução do planejamento orçamentário hospitalar e governamental, pois a ineficiente alocação de recursos conduz a desperdícios (GONZÁLEZ et al., 2019).

Controlar e identificar fatores que afetam o processo de gerenciamento de unidades hospitalares depende da análise dos bancos de dados de saúde. Contudo, esses dados de saúde não são estruturados e são muito complexos. Sua análise requer a utilização de técnicas de aprendizagem de máquina, modelos não lineares e abordagens utilizando múltiplos algoritmos (GIACALONE et al. 2018). As informações de saúde possuem características complexas, dinâmicas e heterogêneas, dificultando a extração de conhecimento utilizando as técnicas tradicionais de análise de dados e estatística. Sobreposta a essa situação, também existe uma lacuna na identificação dos melhores métodos e ferramentas para interpretação dos dados disponíveis (MEHTA; PANDIT, 2018).

Informações redundantes estão armazenadas em algumas situações. As informações de saúde pública no Brasil estão desconectadas em diferentes sistemas, dificultando a sua interpretação quando existe a necessidade de processar registros provenientes de banco de dados distintos (PIRES, 2011).

Outrossim, profissionais de saúde habitualmente não dispõe do conhecimento necessário para conduzir a extração da informação latente em bancos de dados (CHICCO, 2017). Bancos de dados de saúde possuem informações valiosas, mas os profissionais que as utilizam ainda não dominam as ferramentas que permitam a identificação das inter-relações e tendências nesses dados (KUMAR; KHATRI, 2017).

1.2 HIPÓTESE

A construção de modelos de dados, obtidos através da mineração de dados de registros de atendimentos ambulatoriais e de internações hospitalares do sistema público de saúde brasileiro, que permitem a adequada predição da morbidade hospitalar.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo desta pesquisa foi avaliar a predição de morbidade hospitalar, através da aplicação de diferentes métodos de mineração de dados nos registros de procedimentos ambulatoriais e hospitalares, obtidos dos bancos de dados de saúde pública do Brasil.

1.3.2 Objetivos Específicos

Complementarmente, os objetivos específicos foram:

a) indicar um intervalo de tempo adequado para a predição da morbidade hospitalar, executando os processos de mineração de dados propostos nesta pesquisa;
b) identificar o melhor método de mineração de dados, entre aqueles aplicados nos processos de mineração de dados utilizados; e
c) identificar o grupo de morbidades hospitalares que alcança melhor resultado de correlação, quando executados os processos de mineração de dados propostos.

1.4 JUSTIFICATIVA

Para permitir a sustentabilidade e o adequado gerenciamento de sistemas de saúde é necessário minimizar o efeito da aleatoriedade na demanda por serviços, que impacta diretamente na sua eficiência (KUMAR; ANJOMSHOA, 2018). O aprimoramento da capacidade de gestão na aplicação dos recursos de saúde existentes permite a redução da superlotação de unidades hospitalares, a melhora na experiência do paciente e, ao mesmo tempo, a redução dos custos. Um gerenciamento mais eficiente é essencial para reduzir os riscos envolvidos nos cuidados de saúde aos pacientes (GONZÁLEZ et al., 2019).

Bancos de dados dos sistemas de saúde arquivam informações valiosas, que podem ser utilizadas para aprimorar os mecanismos de gestão e melhorar a tomada de decisão de profissionais de saúde. É importante avaliar a possibilidade de prospecção de conhecimento útil a partir dos seus dados.

Modelos epidemiológicos tradicionais possuem limitações quando são aplicados em conjuntos de dados com maior complexidade. Por outro lado, com a mineração de dados – que se apresenta como uma combinação de diferentes disciplinas: Inteligência Artificial, Estatística, Matemática, Sistemas de Banco de Dados – é possível ampliar a capacidade de descobrir padrões, extrair conhecimento e prever resultados de eventos futuros ou desconhecidos (BELLINGER et al., 2017).

Os dados utilizados para disseminação de informação obrigatória de saúde, mormente existentes nos sistemas públicos, ainda possuem a vantagem de estarem armazenados em repositórios abertos e facilmente acessíveis. Do mesmo modo, esses dados não possuem restrição de uso, pois não contêm informações que possam identificar os pacientes; as quais, neste caso, estariam legalmente protegidas e demandariam cuidadosas ponderações éticas para sua utilização (BRASIL, 2010).

Consequentemente, a mineração de dados de saúde é um método promissor para a criação de valor e provisão de introspecções valiosas, sem a sobrecarga de tarefas inerentes às análises estatísticas tradicionais (CHAN et al., 2017). Os modelos preditivos derivados do aprendizado de máquina possuem precisão maior do que o método heurístico. Ainda utilizado, o método heurístico caracteriza-se pela aplicação da experiência humana para o suporte a tomada de decisão (SHIMODA et al., 2018).

A utilização de mineração de dados para a descoberta de padrões desconhecidos e relações importantes em um banco de dados contribui para a análise de uma variedade de dados complexos, permitindo a geração de conhecimento útil (KUNJIR et al., 2017). Para o

gerenciamento adequado de sistemas de saúde, faz-se necessário a disponibilidade de informação confiável, a qual pode ser obtida a partir da mineração de dados. A aplicação de modelos preditivos é possível através do desenvolvimento de sistemas que facilitem o acesso à informação, disponibilizando indicadores personalizados que contribuam para o processo de tomada de decisão (FREIRE et al., 2015).

Através desse conhecimento é possível a predição de eventos relevantes, importante para o aprimoramento da gestão de unidades hospitalares e a correta aplicação dos recursos limitadamente disponíveis. O gerenciamento adequado do conhecimento pode permitir a otimização dos recursos disponíveis, não apenas os financeiros, mas também a disponibilidade de profissionais de saúde, equipamentos de diagnóstico e tratamento.

Portanto, a aplicação de métodos de mineração de dados, com a utilização de algoritmos de aprendizagem de máquina, pode contribuir para o aumento da capacidade de gerenciamento de hospitais públicos e privados. Igualmente, seu emprego em sistemas informatizados tem a capacidade de aprimorar os cuidados de saúde disponibilizados aos pacientes.

1.5 ORGANIZAÇÃO DO TEXTO

Esse capítulo 1 Introdução teve o propósito de contextualizar o problema apresentado, apresentar a hipótese formulada, os objetivos gerais e específicos da pesquisa, bem como a justificativa para o seu desenvolvimento.

O capítulo 2 Mineração Dados esteia os principais processos que foram aplicados na pesquisa e serviram de fundamentação teórica para seu desenvolvimento. Esse capítulo está dividido em 2.1 Classificações de Mineração de Dados, 2.2 Método CRISP-DM, 2.3 Programa Weka, 2.4 Linguagem R, 2.5 Algoritmos para Modelagem de Dados, 2.6 Algoritmos para Seleção de Atributos e 2.7 Métricas de Avaliação.

Outros estudos que detinham alguma correlação com os propósitos ou métodos da presente pesquisa estão apresentados no capítulo 3 Pesquisas Correlatas. Nessa seção, os estudos foram dispostos em ordem cronológica.

Todos os procedimentos executados nessa pesquisa estão apresentados no capítulo 4 Método de Pesquisa, o que permite a replicação dos resultados alcançados. O capítulo está dividido em 4.1 Coleta dos Dado, 4.2 Preparação dos Dados, 4.3 Modelagem e 4.4 Avaliação. Em seguida, os resultados obtidos encontram-se retratados no capítulo 5 Resultados, divididos conforme os intervalos de predição propostos a serem analisados nessa pesquisa.

O capítulo 6 Discussão fundamenta as conclusões que estão no capítulo seguinte e está dividido em 6.1 Mineração dos Dados em Saúde, 6.2 Interpretação dos Resultados, 6.4 Aplicações e 6.5 Desafios para Utilização da Mineração de Dados em Saúde.

2 MINERAÇÃO DE DADOS

Mineração de dados permite a geração de conhecimento em conjunto de dados complexos. Sua aplicação proporciona importantes e valiosas introspecções em problemas multidimensionais, nos quais os métodos estatísticos tradicionais geralmente não alcançam bons resultados (BELLINGER et al., 2017). Deveras, existe uma linha tênue delimitando as diferenças entre as técnicas aplicadas na mineração de dados e na estatística tradicional, pois desempenham o mesmo objetivo: a análise de dados.

Os processos estatísticos podem ser úteis para realização de análises descritivas de dados, mas geralmente não fornecem métodos para realizar modelagens mais avançadas (SHEARER, 2000).

Processos de mineração de dados aplicam análises estatísticas para obtenção de seus resultados. Contudo, trivializando em direção a uma diferença mais marcante, a análise estatística usualmente atém-se a um teste de hipóteses, enquanto o aprendizado de máquina, aplicado na mineração de dados, formula uma generalização para encontrar hipóteses possíveis (WITTEN et al., 2016).

Em Ciência da Computação, a descoberta de conhecimento em banco de dados – ou *knowledge discovery in databases* (KDD) – é definido como o processo, obrigatoriamente envolvendo a concepção de inferências, para identificação em um conjunto de dados de padrões válidos e novos, potencialmente úteis e compreensíveis. O principal propósito do KDD é aplicar técnicas para conceber sentido aos dados analisados, de modo que eles sejam dispostos sinteticamente, permitindo a abstração da informação e gerando conhecimento útil. (FAYYAD et al., 1996).

Processos conduzidos pelo KDD tem objetivo de evidenciar informação desconhecida em um banco de dados, permitindo a geração de novo conhecimento e produzindo uma aplicação prática (KHUMAR; KHATRI, 2017; KUNJIR et al., 2017). É uma área de pesquisa crescente em Ciência da Computação, que permite o aprimoramento dos processos de suporte a decisões, mediada pela análise de grande quantidade de dados e a evidenciação automatizada de padrões (ALMADANI& ALSHAMMARI, 2018).

Conforme sustentado por Fayyad et al. (1996), "*Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.*"

Mineração de dados pode ser aplicada com o propósito de analisar grandes conjuntos de dados, descobrir padrões, extrair conhecimento latente e prever resultados de eventos futuros

ou desconhecidos (BELLINGER et al., 2017). Através dos algoritmos de aprendizado de máquina é possível o desenvolvimento de modelos preditivos, que podem ser usados para prever um intervalo de saídas, como respostas binárias, rótulos categóricos ou valores numéricos (CAMACHO et al., 2018).

A mineração de dados é uma atividade interativa e iterativa (FAYYAD et al., 1996). Através de acertos e erros, paulatinamente os processos envolvidos na sua consecução são aprimorados. Também são constantemente repetidos, buscando-se encontrar a melhor forma de atingir o objetivo proposto. Mesmo porque diferentes conjuntos de dados necessitam de métodos de mineração distintos. Ou ainda, mesmo nas situações em que o conjunto de dados é o mesmo, mas os objetivos são divergentes, os processos de mineração de dados podem precisar ser completamente diferentes.

Não existe um método único. Somente através da compreensão dos dados que são utilizados nos processos, aplicando o conhecimento das características peculiares a cada um dos diferentes algoritmos de mineração de dados, é possível alcançar bons resultados. Todavia, os processos devem seguir um método padronizado, de modo a serem adequadamente documentados, produzindo resultados que possam ser replicados e permitindo a implementação dos modelos em sistemas informatizados.

2.1 CLASSIFICAÇÕES DE MINERAÇÃO DE DADOS

Empregar um enfoque cartesiano para categorizar a mineração de dados é uma tarefa complexa, não apenas porque autores a classificam sob diferentes perspectivas, mas também porque é uma ciência em constante desenvolvimento. No entanto, é importante ressaltar as características dos principais objetivos de mineração de dados, para possibilitar a plena compreensão daqueles propostos nesta pesquisa.

Fayyad et al. (1996), sob o aspecto do objetivo dos processos de mineração de dados, classificaram-na em:

a) métodos preditivos, que são utilizados para predizer valores desconhecidos ou futuros de atributos de interesse, utilizando outros atributos ou campos em banco de dados; e
b) métodos descritivos, os quais concentram-se em encontrar padrões interpretáveis, realizando uma descrição dos dados disponibilizados.

Um conceito aplicado para a detecção de sinais adaptativos e o reconhecimento de padrões em problemas de classificação também é utilizado para dividir a forma como os algoritmos executam a aprendizagem de máquina nos processos de modelagem: aprendizagem supervisionada e aprendizagem não supervisionada (COOPER; COOPER, 1964). Métodos preditivos usualmente utilizam aprendizagem supervisionada, de modo que os descritivos normalmente aplicam aprendizagem não supervisionada.

Aprendizagem supervisionada é executada em conjunto de dados que já foram categorizados, sendo utilizados ao mesmo tempo para a construção dos modelos e para testar as predições realizadas. Já os algoritmos de aprendizagem não supervisionada não possuem dados categorizados e podem ser aplicados para descrever suas características, incluindo proceder o agrupamento dos registros que possuem as mesmas particularidades (COOPER; COOPER, 1964).

Não obstante, essa classificação também foi ampliada em outros tipos de métodos de aprendizagem de máquina. Atualmente, são realizados projetos de mineração de dados através da aprendizagem semi-supervisionada, utilizando conjunto de dados que possuem dados rotulados e não rotulados (HUSSIEN et al., 2017; CAMACHO et al., 2018).

Outro método contemporaneamente aplicado é a aprendizagem por reforço (BIBAULT et al., 2016). Usualmente utilizada em robótica, os algoritmos realizam aprendizagem com interações ambientais, de maneira que através de sua própria experiência constrói uma relação de causa e efeito.

Com ênfase aos propósitos desta pesquisa e seguindo para as subclassificações, a mineração de dados preditiva pode ser categorizada em:

a) problemas de classificação, em que o atributo-classe (que rotula um determinado registro) é uma variável qualitativa (ou categórica) nominal ou ordinal; e
b) problemas de regressão, nos quais o atributo-classe é uma variável quantitativa (ou numérica) discreta ou contínua (FAYYAD et al., 1996; WITTEN et al., 2016).

Considerando os métodos descritivos, a mineração de dados é usualmente subclassificada em associação e agrupamento – ou, como esse último é mais conhecido em inglês, *clustering* (BELLINGER et al., 2017). Mas também, outros propósitos de mineração de dados, como a detecção de desvios em conjunto de dados, são descritos na literatura (WITTEN et al., 2016).

2.2 MÉTODO CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM), em português Processo Padrão Inter-Indústrias para Mineração de Dados, é um método para mineração de dados não-proprietário, documentado e disponível gratuitamente. Foi concebido com objetivo de padronizar e auxiliar a execução de projetos de mineração de dados, disseminando boas práticas e auxiliando a obtenção da maturidade dos processos envolvidos, conforme relatado por Shearer (2000). De acordo com o mesmo autor, "*This model encourages best practices and offers organizations the structure needed to realize better, faster results from data mining.*"

O método CRISP-DM pode ser especializado ou adaptado conforme a necessidade do projeto de mineração de dados. Para o desenvolvimento do processo padrão são apresentadas seis fases:

a) entendimento do negócio;
b) entendimento dos dados;
c) preparação dos dados;
d) modelagem;
e) avaliação; e
f) implementação.

Em cada uma das fases, Shearer (2000) também apresentou diversas etapas necessárias para alcançar os resultados esperados. Desta forma, a seguir são sinteticamente apresentadas informações referentes às etapas e fases da versão 1.0 do CRISP-DM.

A fase entendimento do negócio tem o propósito de prover a compreensão dos objetivos do projeto de mineração de dados. As etapas envolvidas buscam definir adequadamente o problema de mineração de dados e desenvolver o planejamento inicial para alcançar os objetivos. São etapas dessa fase: determinar os objetivos de negócios, avaliar a situação, determinar as metas de mineração de dados e produzir o plano de projeto.

A coleta dos dados introduz a fase de entendimento dos dados, posteriormente seguida pela habituação aos dados e a determinação de sua usabilidade nos processos de mineração. As etapas envolvidas são: coleta de dados iniciais, descrição, exploração e verificação da qualidade dos dados.

Cinco etapas envolvem a fase de preparação dos dados, sendo elas a seleção, limpeza, construção, integração e formatação dos dados. Essa fase também pode ser definida como o

pré-processamento que contiguamente antecede a aplicação dos algoritmos para a modelagem. Seu propósito é transformar o conjunto de dados de maneira a tornar a mineração mais eficiente. Essa etapa é fundamental para alcançar os objetivos do projeto, tornando-se, em muitos deles, tão ou mais importante do que a própria execução dos algoritmos de modelagem.

Na fase de modelagem são aplicados os algoritmos de mineração de dados, precedidos de sua parametrização e calibração. Diferentes técnicas e algoritmos são empregados de acordo com o conjunto de dados a ser analisado. As etapas de modelagem incluem seleção da técnica de modelagem, geração do desenho de teste, criação e análise dos modelos. A análise do modelo deve ocorrer em dados com a mesma estrutura, mas preferencialmente não utilizados na etapa de modelagem. Os resultados da mineração de dados são avaliados através de diferentes métricas, de acordo com o objetivo proposto, o resultado esperado e o conjunto de dados utilizado.

Na fase avaliação, as etapas envolvidas são: avaliação dos resultados, revisão do processo e determinação dos próximos passos. Um possível passo após a fase de avaliação seria a conclusão do projeto, conduzindo-o para a última fase denominada implementação. Existe ainda a possibilidade de reiniciar etapas anteriores, objetivando iterações nos processos de mineração, com o propósito de aprimorá-los.

A fase de implementação caracteriza-se pela aplicação dos modelos obtidos em aplicações profícuas, com a utilização do conhecimento evidenciado durante todas as fases do projeto, não apenas com o resultado da mineração de dados. Em alguns casos, a fase de implementação pode estar limitada a produção de um relatório, mas em outros envolve a aplicação dos modelos de dados em algoritmos de diferentes sistemas informatizados. As etapas envolvidas na fase de implementação podem incluir: planejar a implementação, planejar o monitoramento e a manutenção, produção do relatório final e revisão do projeto.

2.3 PROGRAMA WEKA

O programa *Waikato Environment for Knowledge Analysis* (Weka), escrito em linguagem JAVA e produzido pela Universidade de Waikato, na Nova Zelândia, é utilizado para mineração de dados. O programa é distribuído sob uma licença GNU General Public License e consiste em uma coleção de algoritmos de aprendizado de máquina para a realização de processos de mineração de dados, incluindo preparação, classificação, regressão, clusterização e regras de associação, bem como diversas interfaces para a visualização dos resultados e para sua análise estatística. (WITTEN et al., 2016).

O Weka agrega diversos ambientes que permitem a realização de pesquisas utilizando o estado-da-arte em mineração de dados, mantendo-se constantemente atualizado (NAVAZ et al., 2016). O programa utiliza a implementação de diversos algoritmos de aprendizagem de máquina, empregando conceitos de Inteligência Artificial. Permitindo, desta forma, a aplicação de técnicas que já foram amplamente descritas na literatura e aplicadas em diversos projetos de mineração de dados.

2.4 LINGUAGEM R

R pode ser referido como uma linguagem de programação e um ambiente de desenvolvimento integrado, que permite a realização de computação estatística e gráfica. Considerado como uma implementação da linguagem S, é um programa livre distribuído sob os termos da licença GNU General Public License (BECKER et al., 1988).

A linguagem R contém um conjunto de recursos para manipulação de dados, realização de cálculos e exibição gráfica, agregando uma coleção integrada de ferramentas para análise e modelagem de dados, sendo considerada uma linguagem de programação simples e eficaz. Da mesma forma, oferece diversos pacotes com funções que simplificam a execução de análises estatísticas (VENABLES; SMITH, 2009).

2.5 ALGORITMOS PARA MODELAGEM DE DADOS

Mineração de dados preditiva supervisionada em conjunto de dados com atributos numéricos, também descrita sob a forma de um problema de regressão, é habitualmente realizada utilizando árvores de regressão, máquinas de vetores de suportes e redes neurais artificiais (BELLINGER et al., 2017; RICHTER; KHOSHGOFTAAR, 2018). Regressão linear também é uma técnica natural a ser considerada em conjunto de dados numéricos, mas pode ser classificada apenas como uma técnica estatística. Por último, técnicas de aprendizado baseado em instâncias, que utilizam o hiperespaço para modelagem dos dados, são naturalmente bem-sucedidas com atributos numéricos, em razão dos cálculos baseados em distância dos elementos (WITTEN et al., 2016).

Adiante estão descritos algoritmos estudados e desenvolvidos por diversos pesquisadores de Ciência da Computação, alicerçados em cálculos e provas dos fundamentos de Matemática da Computação. Apesar da pretensão inicial de demonstrar esses fundamentos

dos principais algoritmos, o intuito foi explorar a compreensão dos motivos pelos quais os resultados preditivos são alcançados, e não como.

Uma das principais questões da mineração de dados de saúde é exatamente a necessidade de entendimento mútuo das ferramentas disponíveis para mineração, assim como o contexto e significado das informações extraídas dos dados analisados. Apenas com um profundo entendimento das relações existentes nos bancos de dados é possível compreender os resultados encontrados, utilizando as técnicas de análise de dados; produzindo, desta forma, o conhecimento necessário para a alcançar a inteligência clínica esperada e resultando em uma aplicação profícua em saúde (GIACALONE et al., 2018). Mas pouca relevância teria em aprofundar as discussões tanto em relação aos fundamentos dos algoritmos, como também quanto aos aspectos clínicos e fisiológicos das condições de saúde representadas nos conjuntos de dados.

Não obstante, é importante para profissionais de saúde entenderem os aspectos mais abstratos dos procedimentos de mineração de dados, para compreensão dos motivos pelos quais essas ferramentas podem aprimorar seu poder de decisão. Da mesma forma, os significados das informações de saúde, sobremaneira quanto ao aspecto epidemiológico dos dados existentes, podem auxiliar os profissionais de Ciência da Computação – e suas subáreas – no desenvolvimento de processos mais eficientes e eficazes.

Conseguintemente, a seguir, conceitos amplamente estruturados dos algoritmos que foram utilizados no processo de modelagem da mineração de dados preditiva realizada, que executam processos de aprendizagem supervisionada, estão descritos.

Árvores de decisão utilizam uma abordagem do tipo “dividir e conquistar”, realizando a divisão do conjunto de dados por meio do cálculo do ganho de informação para escolha dos atributos mais significativos para sua estruturação, alocando sucessivamente os atributos em nós condicionais das divisões construída. O algoritmo M5, originalmente desenvolvido por Quinlan (1992), constrói uma árvore de regressão, utilizando modelos lineares multivariados; de modo que os modelos se tornem funções lineares separadas em diversos componentes. Em cada nó da árvore construída são realizados testes para definir as divisões que maximizem a redução do erro nas distintas funções lineares, bem como possibilitem o maior ganho de informação possível. Após a formação da árvore de decisão, uma fórmula de regressão linear é construída utilizando apenas o subconjunto de dados delimitado.

Algoritmos executados através dos conceitos de aprendizagem em conjunta – ou, em inglês, *ensemble learning* – podem obter melhores resultados do que quando utilizados individualmente. O algoritmo Random Forest foi descrito por Breiman em 2001 e constitui-se

da combinação de árvores de decisão para a modelagem de dados. Random Forest constrói as árvores de decisão utilizando subconjuntos de dados de forma aleatória, desconsiderando a variância entre as amostras. O grande número de árvores de decisão construídas conduz a uma pequena variância entre os subconjuntos, obtendo melhores resultados do que outros métodos de *ensemble learning*, ao produzir uma combinação mais aleatória possível. Para problemas de regressão, o algoritmo utiliza os valores numéricos, ao invés de classes, nos vetores do conjunto de árvores; atribuindo a média aritmética dos valores de predição em cada árvore como resultado.

Algoritmos baseados em aprendizagem por instâncias não produzem modelos, apenas armazenam os registros do subconjunto de dados de treinamento. O processo de predição somente é realizado quando uma nova instância precisa ser classificada. Aha et al., em 1991, descreveram o algoritmo para "aprendizagem por exemplos", que aborda o hiperespaço do conjunto de dados de maneira incremental, assumindo que instâncias similares possuem a mesma classificação. O aprendizado por instâncias diminui a custos computacionais incorridos na abstração de conceitos que podem não ser utilizados para a classificação de uma nova instância; desta forma, aumentando a taxa de aprendizado. O algoritmo procura os *n* vizinhos mais próximos – ou, em inglês, *k-nearest neighbour* (kNN) – no hiperespaço para realizar a predição, utilizando para os cálculos, por exemplo, a distância euclidiana.

As máquinas de vetores de suporte – em inglês, *support vector machines* (SVM) foram desenvolvidas através dos fundamentos das pesquisas de Vapnik. Em seus sucessivos trabalhos, o autor buscava um melhor desempenho na generalização de classificadores, maximizando a margem do limite de decisão no hiperespaço através de uma combinação linear de vetores de suporte mais próximos desse limite. Seus estudos foram iniciados na busca de um aprimoramento de outros algoritmos, como o Perceptron e a função de base radial (BOSER et al. 1992). Conceitos de delimitação do hiperespaço já existiam, mas ao inverter a ordem de operações para construir a função de decisão – primeiro, comparando dois vetores no hiperespaço e, em seguida, realizando a transformação não-linear do valor de resultado – possibilitou a criação de superfícies de decisão polinomiais. Esse algoritmo de aprendizagem de máquina foi então denominado de rede de vetores de suporte (CORTES; VAPNIK, 1995).

Posteriormente, Vapnik consolidou seus fundamentos, aprimorando a forma de cálculo para a transformação do hiperespaço não-linear, através da aplicação de diferentes funções nucleares (*kernel*). A utilização desse processo possibilita uma melhor distribuição linear do conjunto de dados para a delimitação de hiperespaço e, em seguida, a definição dos vetores de suporte e da margem máxima do limite de decisão. "*Using the method of the optimal separating*

hyperplane we construct a new class of learning machines for estimating indicator functions, the socalled support vector machines (...)" (VAPNIK, 1998).

Em 1997, foi desenvolvido uma versão de SVM para ser aplicada em problemas de regressão. As máquinas de vetores de suporte aplicadas para a regressão possuem vantagens em hiperespaços de alta dimensionalidade, em razão da forma como os cálculos de otimização são realizados. Desta forma, o algoritmo possui grande utilidade em conjunto de dados que contenham uma dimensionalidade de representação de espaço maior do que a do número de exemplos (DRUCKER et al., 1997).

Em meados do século XX, pesquisadores buscavam desenvolver um modelo de dados baseado no funcionamento do cérebro e seus neurônios. Em 1958, Rosenblatt teorizou o algoritmo Perceptron, utilizando uma abordagem estatística quantitativa para explicar a organização dos sistemas cognitivos, que possuía a perspectiva de replicar os processos de funcionamento de um neurônio. As redes neurais artificiais são usualmente compostas por conexões e agrupamentos do algoritmo Perceptron. Uma das implementações de rede neural artificial multicamadas é conhecida como Multilayer Perceptron, que conecta os nós da camada de entrada e os da camada de saída através de uma ou mais camadas intermediárias ou ocultas.

O processo de treinamento do Multilayer Perceptron é realizado através do algoritmo Backpropagation, concebido por Werbos (1974) em sua tese de doutorado. O algoritmo permitiu o aprimoramento das redes neurais artificiais, ao realizar cálculos para a modificação dos pesos das conexões entre os Perceptrons e, em seguida, mensurar a contribuição de cada um na obtenção do resultado correto durante as iterações ocorridas na modelagem dos dados. Para tanto, é utilizada uma estratégia de otimização matemática, mormente aplicada para ajustar os parâmetros de uma determinada função, denominada descida do gradiente. O algoritmo Backpropagation, ao aprimorar o funcionamento das redes neurais artificiais, viabilizou a continuidade das pesquisas sobre esse assunto.

2.6 ALGORITMOS PARA SELEÇÃO DE ATRIBUTOS

Limitar o número de atributos utilizados na fase de modelagem confere vantagens no processo de mineração de dados, como, por exemplo, a minimização da influência de atributos com menor relevância nos modelos e daqueles com valores repetidos. Outro benefício é o menor comprometimento da capacidade de processamento e de memória dos computadores empregados nos processos de mineração (WITTEN et al., 2016).

A utilização de algoritmos de seleção de atributos é particularmente importante em conjunto de dados de alta dimensionalidade. Comparando com outras técnicas de redução de dimensionalidade, como a análise de componentes principais, as técnicas de seleção não modificam a representação original dos atributos, induzindo a uma melhor interpretabilidade dos resultados obtidos (SAEYS et al., 2007).

Ao aproveitar as proficuidades da seleção dos atributos, muitos projetos de mineração de dados obtêm resultados que superam aqueles obtidos sem a sua aplicação. Constituindo seus algoritmos, dessa forma, ferramenta tão importante quanto aqueles empregados nos processos de modelagem (HUSSIEN et al., 2017).

Entre as principais abordagens, que são utilizadas pelos algoritmos que executam a seleção de atributos, está a seleção baseada no melhor resultado obtido em repetidos processos de validação cruzada; que testam o melhor subconjunto de dados, aplicando o próprio algoritmo que será utilizado na modelagem. Outra utiliza-se de cálculos para selecionar os atributos que possuem maior associação com a classe a que pertence o registro, normalmente aplicando o ganho de informação. Os processos que executam validação cruzada na matriz de atributos, buscando encontrar o conjunto que proporciona os melhores resultados, possuem alto custo computacional, pois realizam uma extensa varredura nas possíveis combinações. Contudo, em conjunto de dados menores, esses processos podem ser muito eficientes, notadamente conduzindo a resultados superiores (WITTEN et al., 2016).

Nas abordagens que calculam o ganho de informação, é construída uma lista classificada dos melhores atributos. Desta forma, é mensurada a capacidade que cada atributo, ou o conjunto desses, pode contribuir no aprimoramento dos modelos. Contudo, esse aprimoramento só pode ser quantificado após a execução dos procedimentos de teste dos modelos (TAZIN et al., 2016; TESFAYE et al. 2017).

Considerando os propósitos desta pesquisa, o algoritmo avaliador de subconjunto de atributos baseado em correlação – *correlation-based attribute subset evaluator* – sustenta-se na hipótese de que atributos possuem valores intimamente correlacionados aos valores da classe (MLAKAR et al., 2018). Os subconjuntos são avaliados pela capacidade preditiva de cada atributo separadamente e seu grau de inter-redundância. O algoritmo pode ser executado em conjunto com um processo concomitante de seleção direta e pesquisa por eliminação retrógrada de atributos, aplicado pelo algoritmo Greedy; o qual também é capaz de inferir a redundância dos atributos (WITTEN et al., 2016).

2.7 MÉTRICAS DE AVALIAÇÃO

No processo de treinamento da modelagem de dados, métricas de avaliação são utilizadas para otimizar o algoritmo de classificação. Nos processos de testes do modelo, as métricas têm o propósito de aferir a eficácia do classificador produzido (HOSSIN; SULAIMAN, 2015).

Diferentes métricas podem ser aplicadas para avaliação dos modelos construídos pelos algoritmos de aprendizagem de máquina. Problemas de classificação ou de regressão podem compartilhar algumas mesmas métricas, como também deter aquelas que são específicas a cada; ainda, em uma visão mais subjetiva, conjunto de dados com características diferentes, ou objetivos distintos do projeto de mineração de dados, requerem abordagens e métricas de avaliação específicas.

Para que uma métrica de avaliação seja considerada adequada, é necessário possuir capacidade de avaliação do impacto do desempenho, especificamente no domínio que está sendo avaliado (WAGSTAFF, 2012; SHICKEL et al., 2018).

Contudo, em mineração de dados uma situação é única: o conjunto de dados utilizado para a modelagem ou treinamento dos algoritmos deve ser diferente do aplicado para a avaliação do modelo obtido. Essa é a regra. Avaliar um modelo nos mesmos dados que foram utilizados para treinamento conduz a uma superestimação da eficiência dos modelos.

Alguns métodos utilizados para a correta avaliação do modelo de mineração de dados são: a divisão em conjunto de dados de treinamento e conjunto de dados de teste; e o processo de avaliação cruzada – *cross-validation*, em inglês. A divisão em treinamento e testes, como é inerente à sua denominação, fraciona o conjunto de dados para a execução dos procedimentos de modelagem e avaliação separadamente. O processo *cross-validation*, apesar de um pouco mais complexo, permite que, mesmo em um conjunto de dados desbalanceado ou com dados faltantes, os resultados sejam uma representação mais fidedigna das características do modelo. Utilizando um exemplo para sua definição, um procedimento de *10-fold cross-validation* divide o conjunto de dados em dez fragmentos, executa a modelagem em nove desses e avalia os resultados em apenas um. O processo é repetido até que os procedimentos de avaliação do modelo sejam aplicados em todos os fragmentos, utilizando os modelos construídos em dez diferentes conjuntos de dados de treinamento. O resultado final é a média das métricas de avaliação do modelo em cada um dos fragmentos (WITTEN et al., 2016).

Os modelos matemáticos dos algoritmos de aprendizagem de máquina usualmente buscam sua otimização através da minimização do erro de predição. Consequentemente,

métricas que utilizam o erro nas variáveis de cálculos são amplamente aplicadas em mineração de dados. Raiz quadrada do erro quadrático médio – *root mean squared error* (RMSE) – é uma métrica usualmente aplicada para avaliação de modelos de mineração de dados, pois permite a definição da diferença entre o valor previsto e o valor observado. Outra métrica que utiliza cálculos baseados nos erros dos modelos é a média do erro absoluto – *mean absolute error* (MAE) (PELÁNEK, 2015).

Ao invés dos valores absolutos, também podem ser utilizados valores relativos para mensuração do erro dos modelos. A média da porcentagem absoluta do erro – em inglês, *mean absolute percentage error* (MAPE) – calculam o erro relativo dos modelos, permitindo a comparação de resultados de diferentes conjuntos de dados (SANTOS; CARVALHO, 2018).

Algumas métricas utilizam cálculos que permitem uma avaliação mais ampla do modelo, ao mensurar sua capacidade de predição. Acurácia – *accuracy* em inglês – é uma métrica amplamente utilizada, que considera a proporção de instâncias classificadas corretamente. Estatística Kappa – em inglês, *Kappa Statistics* – mede o desempenho entre dois conjuntos de dados classificados, possibilitando a comparação do modelo obtido com um aleatório; sendo mais robusto do que uma simples proporção de desempenho (KHUMAR; KHATRI, 2017; BAŞAR; AKAN, 2018).

Existe uma forma de interpretar qualitativamente os resultados de testes de um modelo de mineração de dados, através de métricas que utilizam resultados positivos e negativos em um problema de classificação (PELÁNEK, 2015). Algumas delas também são aplicadas em estudos epidemiológicos e consideram a ocorrências de casos positivos e negativos de doenças. Para os cálculos, são utilizadas as incidências de casos positivos verdadeiros e positivos falsos, assim como os negativos verdadeiros e negativos falsos.

Sensibilidade (*sensitivity*) é a proporção de padrões positivos que são classificados corretamente. Já especificidade – ou *specificity* – é a proporção de casos padrões negativos que são adequadamente classificados. A precisão (*precision*) é usada para medir as ocorrências positivas que são corretamente classificadas em razão de todas as predições positivas. E r*ecall* é utilizado para medir as ocorrências positivas, em razão de todas as predições realizadas corretamente (HOSSIN; SULAIMAN, 2015; GUPTA et al., 2017; BAŞAR; AKAN, 2018).

Medida F – também tratada como *F-measure* ou *F1 score* – pode ser considerada uma média harmônica de *precision* e *recall*. Assim, a utilização da métrica *F-measure* permite uma avaliação sistêmica dos resultados obtidos (PELÁNEK, 2015; BAŞAR; AKAN, 2018).

A curva característica de operação do receptor – em inglês, r*eceiver operating characteristic curve* (ROC) – é formada por um gráfico cartesiano com os dados do índice de

positivos verdadeiros (ou sensibilidade), no eixo vertical, e do índice de falsos positivos (ou 1 - especificidade), no eixo horizontal (MIOTTO et al., 2016; CIRKOVIC et al., 2017; WIRATMADJA et al., 2018). A ROC representa o desempenho de um classificador sem considerar a distribuição de uma ou outra classe, sendo particularmente útil em conjuntos de dados desbalanceados. Ao calcular a área sob a curva – *area under the curve* (AUC) – é produzida uma métrica de avaliação que não depende do limiar de discriminação, possibilitando a análise sintética dos resultados de classificação (MIOTTO et al., 2016; PELÁNEK, 2015; WITTEN et al., 2016; TAZIN et al. 2016; RICHTER; KHOSHGOFTAAR, 2018).

Não obstante, com exceção dos métricas baseadas no erro do modelo, essas até aqui apresentadas não são aplicáveis em problemas de regressão. *Accuracy*, estatística Kappa, sensibilidade, especificidade, AUC são métricas utilizadas exclusivamente para problemas de classificação.

Muitos estudos que aplicam mineração em conjunto de dados com classes de valores numéricos usualmente utilizam a RMSE para avaliação dos modelos (PELÁNEK, 2015). Contudo, métricas que representam os erros dos modelos podem ser sensíveis às diferentes grandezas dos atributos e das classes (HOSSIN; SULAIMAN, 2015; WITTEN, 2016). Ou seja, aplicadas na avaliação de modelos em conjunto de dados diferentes produzem resultados com escalas incompatíveis, desta forma, inviabilizando a comparação estatística.

Assim, podem ser utilizadas métricas de avaliação como o coeficiente de correlação de postos de Spearman, o coeficiente de correlação τ de Kendall e o coeficiente de correlação produto-momento (ou ρ de Pearson) (ALSHAHRANI et al., 2018). Entre essas, uma métrica adequada para analisar a correlação entre diferentes variáveis é o coeficiente de correlação ρ Pearson (SACRAMENTO, 2014; SANEJA; RANI, 2019).

Pois, de maneira diferente, os coeficientes de correlação – que medem a relação estatística entre os valores previstos e os valores reais – são independentes da escala do conjunto de dados. Consequentemente, essa é uma característica importante a ser considerada nos projetos de mineração de dados que envolvam conjunto de dados diferentes.

Na avaliação de modelos em problemas de regressão, o coeficiente de correlação produto-momento ou ρ de Pearson é calculado através da fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

, onde *x* é o valor previsto pelo modelo e *y* o valor real da classe no conjunto de dados de teste. Os resultados variam entre -1 (um negativo), quando possuem correlação negativa perfeita, e 1 (um), quando possuem correlação positiva perfeita. E quanto mais próximo do zero pior o resultado obtido pelo modelo avaliado (WITTEN et al., 2016). Consequentemente, a percepção do significado do valor dos resultados dos coeficientes de correlação é intuitiva.

Outrossim, para a classificação do coeficiente ρ Pearson, com objetivo de categorizar o grau de correlação alcançado, Evans (1996) sugeriu a seguinte divisão:

a) entre 0 e |0,19|: correlação muito fraca;
b) entre |0,20| e |0,39|: correlação fraca;
c) entre |0,40| e |0,59|: correlação moderada;
d) entre |0,60| e |0,79|: correlação forte; e
e) entre |0,80| e |1|: correlação muito forte.

Aqui, propositadamente foram apresentadas as nomenclaturas das métricas em inglês e português, pois suas traduções podem variar em diferentes idiomas. Sendo possível, desta forma, que ficassem descaracterizadas.

3 PESQUISAS CORRELATAS

Vianna et al. (2010), integraram informações de diferentes sistemas de informação de saúde pública do Brasil, com objetivo de demonstrar a viabilidade da execução de mineração nesses bancos de dados. Durante a fase de preparação dos dados, incluíram um atributo ao banco de dados, com as informações da Classificação Internacional de Doenças (CID-10) dos registros. Utilizaram uma árvore de decisão – uma implementação do algoritmo C4.5 no programa Weka – para a criação de regras acerca do perfil de mortalidade infantil. Ao final, concluíram que a utilização de mineração de dados poderia gerar importantes conclusões em saúde pública.

Em 2011, Santos e Gutierrez desenvolveram uma plataforma de armazenagem (*data warehouse*) para os bancos de dados de saúde pública do Brasil, com objetivo de realizar extração de informação analítica. Foram utilizadas técnicas de produção de informação gerencial e descoberta de conhecimento, respectivamente denominadas On-line Analytical Processing (OLAP) e On-line Analytical Mining (OLAM), incorporadas à própria plataforma desenvolvida, que utilizam métodos de mineração de dados. Os testes resultaram em diversos níveis de correlação dos dados, possibilitando concluir sobre a existência da capacidade de obtenção de informação útil à gestão de saúde. Os autores afirmaram que havia a necessidade de inclusão de outros métodos de modelagem e técnicas de mineração de dados.

Também foi possível constatar a inclusão do CID-10, como novo atributo em um conjunto de dados para realização de mineração, em um artigo publicado em 2011 por Barros et al. Os autores não executaram processos de modelagem, pois o objetivo era apenas demonstrar um método estruturado de preparação de dados para mineração.

Em 2011, Pires ampliou a pesquisa no ambiente desenvolvido por Santos e Gutierrez, investigando processos mais elaborados para preparação, correlação (*record linkage*), padronização e armazenamento dos dados, aplicando também as técnicas de análises OLAP e OLAM em sua tese de doutorado. O desenvolvimento da pesquisa utilizou, além dos dados provenientes dos registros de saúde pública, registro de atendimentos de um hospital universitário. Foram aplicados métodos para correlação de variáveis, ampliando a capacidade de análise da informação. O pesquisador afirmou que continuaria a desenvolver técnicas de mineração para análise dos dados e esperava que seu trabalho estimulasse pesquisas que utilizassem técnicas de relacionamento de registros, com objetivo de obtenção de informações epidemiológicas da base de dados de saúde pública do Brasil.

Freire et al. (2015) utilizaram o *software* Pentaho para avaliar a ligação entre diferentes bancos de dados de saúde pública do Brasil. A correlação entre dados referentes a registros hospitalares e ambulatoriais de atendimento oncológico foram analisados, assim como dados sobre a mortalidade. As análises tinham como objetivo a construção de um *data warehouse*, que considerasse as correlações entre as informações dos diferentes bancos de dados. O *data warehouse* construído possibilitaria o gerenciamento mais eficiente de informações de pacientes sob tratamento oncológico, incluindo a produção de relatórios com informações relevantes conectadas. Eles conceberam que seria necessário o desenvolvimento de outras estratégias para aprimorar a eficiência dos processos de correlação de informações.

Como propósito secundário em uma pesquisa que desenvolveu modelos quantitativos através de registros eletrônicos de saúde de pacientes, que seriam utilizados para prever seu estado de saúde, Miotto et al. (2016) utilizaram a frequência da ocorrência de um teste de laboratório para determinar padrões nos registros de pacientes. O objetivo era a predição de ocorrências de doenças, sem que houvesse a necessidade de considerar o resultado desses mesmos exames. Os autores afirmaram que incluir os valores dos resultados do teste de laboratório poderia melhorar o desempenho preditivo, mas esses dados não seriam simples de processar em larga escala, pois poderiam estar disponíveis apenas como sinalizadores de texto nos registros eletrônicos de saúde, como também em diferentes unidades de medida.

Em 2017, Baurin et al. apresentaram informações sobre sua pesquisa acerca da mineração em bancos de dados de informações hospitalares, com objetivo de prover informação sobre o impacto das doenças na infraestrutura de saúde. Utilizaram dados do Sistema de Informação Hospitalar (SIH) do sistema público de saúde do Brasil. Os pesquisadores relataram que o banco de dados possuía 99% (noventa e nove por cento) dos registros adequados; a plataforma utilizada permitia a exploração dos bancos de dados massivos; e a mineração de dados detinha a capacidade de prover introspecções úteis sobre a infraestrutura de saúde pública do Brasil.

A habilidade para entender, prever e prevenir a mortalidade infantil foi entendida como importante estratégia para melhorar a qualidade de saúde por Santos e Carvalho. Em um artigo em 2018, os autores aplicaram mineração em diferentes bancos de dados, do repositório do sistema de saúde pública do Brasil, com o propósito de avaliar a capacidade do modelo em predizer a mortalidade infantil. Para tanto, utilizaram algoritmos de regressão linear, redes neurais artificiais e máquinas de vetores de suporte, em um problema de regressão. Dos resultados obtidos, concluíram que as redes neurais artificiais necessitavam de ajustes dos parâmetros refinados para aumentar sua precisão; mas que as máquinas de vetores de suporte

também se mostraram efetiva, sem a necessidade de ajuste de muitos parâmetros para obtenção de previsões precisas.

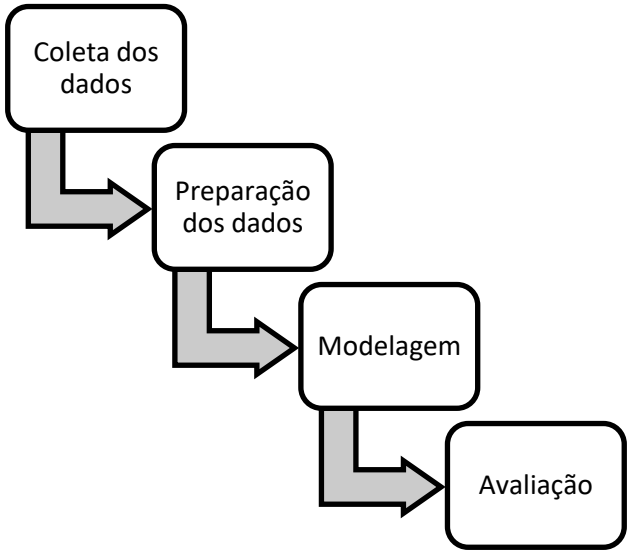
Conseguintemente, a presente pesquisa diferencia-se dos trabalhos anteriores por realizar uma mineração de dados preditiva, com objetivo de produzir um modelo de dados, fundamentada na correlação existente entre procedimentos de diagnóstico em laboratório clínico e morbidades assentadas em internações hospitalares. Tanto quanto se sabe, até o presente momento não existem estudos publicados que aplicaram a mineração de dados com o mesmo propósito.

4 MÉTODO DE PESQUISA

Foi realizada uma pesquisa explicativa quantitativa aplicada, estruturada através de um método hipotético-dedutivo para realização de um estudo experimental. Processos de mineração de dados foram executados em 11 (onze) anos de registros de informações, compreendidos entre 2008 e 2018, referentes a atendimentos ambulatoriais e hospitalares do sistema de saúde pública do Brasil.

Os processos de mineração de dados foram executados utilizando abordagens fundamentadas em algumas fases e etapas do CRISP-DM, conforme Figura 1. Foi realizada uma mineração de dados preditiva, com aplicação de algoritmos que realizam aprendizagem supervisionada, para a modelagem de um problema de regressão.

Figura 1 - Processos de mineração de dados executados.



Fonte: Elaborado pelo autor (2019).

4.1 COLETA DOS DADOS

Os arquivos disseminàveis utilizados nos processos de mineração de dados foram adquiridos dos repositórios do Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS), vinculado à Secretaria de Gestão Estratégica e Participativa do Ministério da Saúde.

Os arquivos encontravam-se em formato de banco de dados compactados do dBase, que foram interpretados e tabulados pelo programa Tabulador de Dados para Windows (TABWIN), versão 4.1.4, disponibilizado pelo DATASUS. O programa consiste em um conjunto de

ferramentas para tabulação e apresentação das informações contidas nos arquivos, que utiliza uma interface gráfica (BRASIL, 2018a).

Também foram utilizados os arquivos das competências de três meses subsequentes ao período proposto. Esse processo teve objetivo de prover dados que eventualmente foram posteriormente apresentados aos sistemas de registro do DATASUS, mas que se referem ao período de avaliação da pesquisa. Uma situação usual nesses sistemas (BRASIL, 2010).

Para a coleta de dados, os arquivos disseminâveis do Sistema de Informações Ambulatoriais do SUS (SIASUS), que são tipificados como "Produção Ambulatorial" (PA), foram obtidos dos repositórios. Os dados eram referentes a procedimentos ambulatoriais, obtidos através dos seguintes instrumentos de registro:

a) Autorização de Procedimentos Ambulatoriais de Alta Complexidade (APAC);
b) Boletim de Produção Ambulatorial Consolidado (BPA-C);
c) Boletim de Produção Ambulatorial Individualizado (BPA-I);
d) Registro das Ações Ambulatoriais de Saúde - Atenção Domiciliar (RASS-AD); e
e) Registro das Ações Ambulatoriais de Saúde - Atenção Psicossocial (RASS-PSI) (BRASIL, 2018b).

Da mesma forma, foram obtidos os arquivos disseminâveis do Sistema de Informações Hospitalares do SUS (SIHSUS), que são tipificados como "AIH Reduzidas" (RD). Esses arquivos continham os registros de Autorização de Internação Hospitalar (AIH) (BRASIL, 2018c).

Um dos arquivos disseminâveis adquiridos do repositório do DATASUS (RDAC0909.dbc), referente ao processamento de informações hospitalares da competência do mês de setembro de 2009 no Estado do Acre, estava corrompido e foi desconsiderado nos procedimentos realizados.

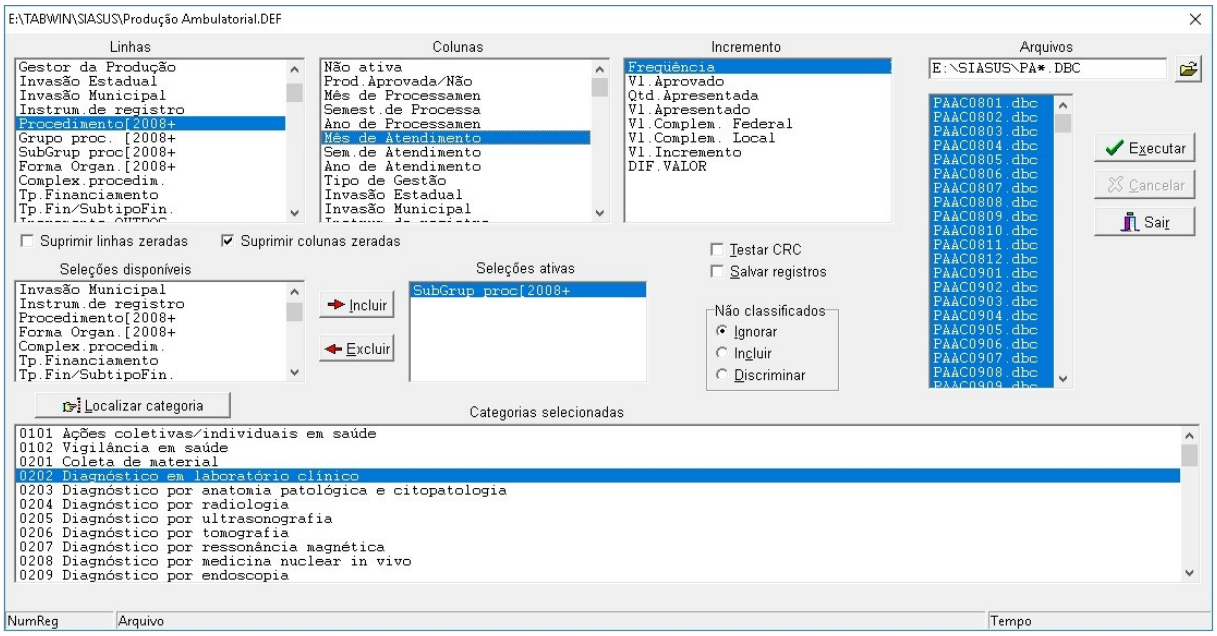
4.2 PREPARAÇÃO DOS DADOS

Os arquivos disseminâveis dos sistemas SIASUS e SIHSUS foram processados através do programa TABWIN.

As frequências de procedimentos ambulatoriais de diagnóstico em laboratório clínico (subgrupo 02.02 no DATASUS) por mês de atendimento foram tabuladas dos registros dos arquivos do SIASUS, conforme Figura 2. Foram englobados dados registrados de todo o Brasil,

que posteriormente foram importados para uma planilha eletrônica, no formato padrão c*omma-separated values* (CSV), através do TABWIN.

Figura 2 - Tela da tabulação executada nos arquivos do SIASUS.



Fonte: Tabulador de Dados para Windows (2019).

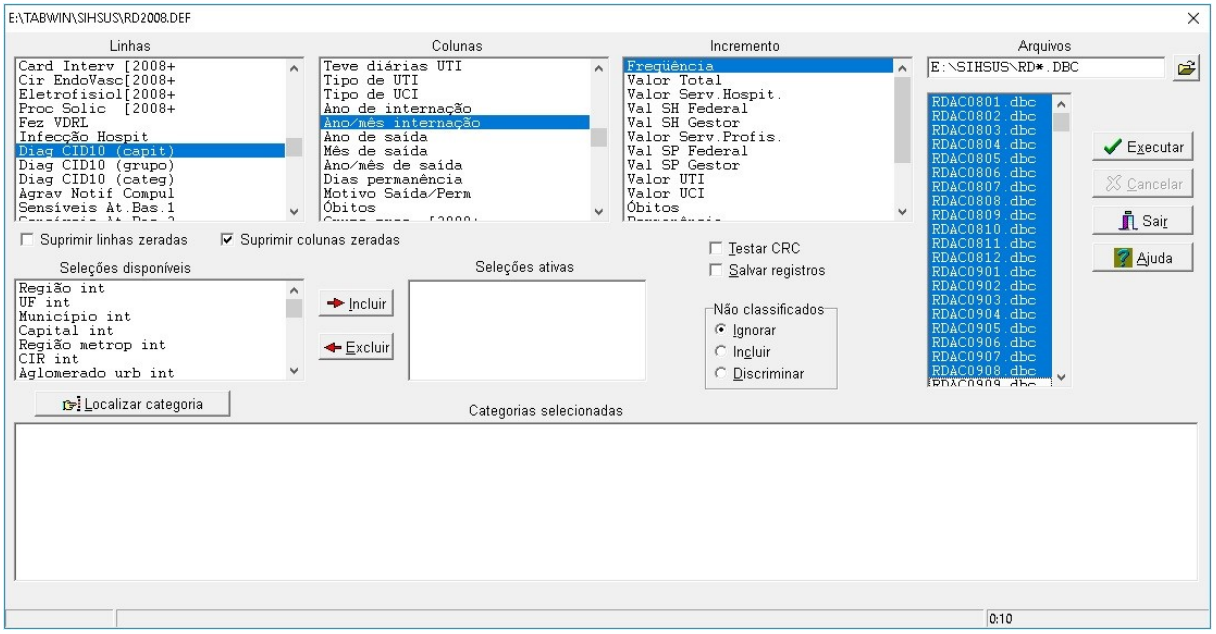
Da igual maneira, foram tabulados os arquivos disseminàveis do SIHSUS em uma tabela contendo a frequência das morbidades registradas nos atendimentos hospitalares, considerando os capítulos da Classificação Internacional de Doenças (CID-10), por mês de internação. Também foram abrangidos os dados registrados de todo o país, os quais foram igualmente importados em formato padrão CSV. A tela do processo de execução da tabulação dos registros de atendimentos hospitalares está ilustrada na Figura 3.

Em seguida, as planilhas eletrônicas resultantes contendo uma série de dados, foram ajustadas de maneira que cada mês de atendimento ou internação – de acordo com a fonte dos dados – fossem instâncias (ou linhas) dos atributos (ou colunas). Cada atributo correspondeu às quantidades da ocorrência de cada procedimento ambulatorial ou de cada morbidade hospitalar. Foram considerados apenas os dados do período proposto da pesquisa, entre os anos de 2008 e 2018, mesmo que o processo de tabulação pudesse gerar outros períodos.

Assim, na tabela de procedimentos ambulatoriais foi incluída, ao final, uma única coluna com cada morbidade hospitalar individualmente. Esse processo foi realizado repetidas vezes, resultando, desta forma, 22 planilhas eletrônicas diferentes, uma para cada capítulo do

CID-10. Reitera-se que cada planilha continha também exatamente o mesmo conjunto de dados de procedimentos ambulatoriais, conforme exemplificado na Figura 4.

Figura 3 - Tela da tabulação executada nos arquivos do SIHSUS.

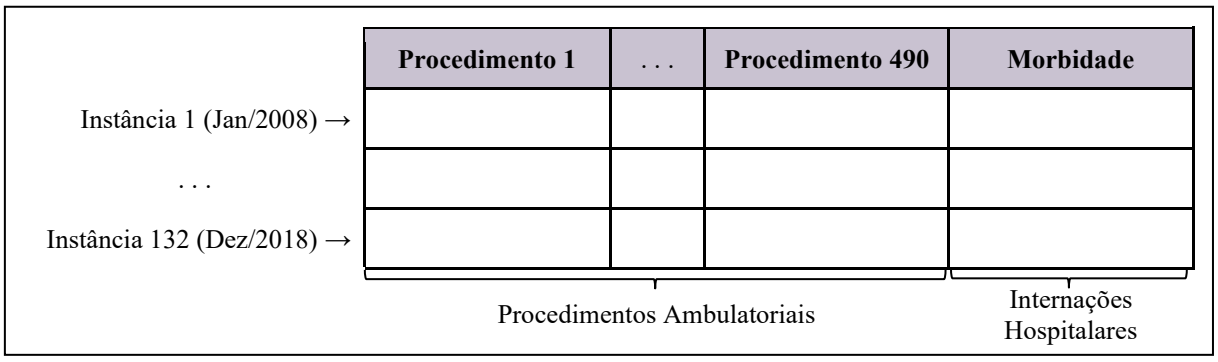


Fonte: Tabulador de Dados para Windows (2019).

Tanto nas etapas de preparação dos dados, quanto na fase de modelagem, foi utilizado o programa Weka, que possibilitou a adequada importação das planilhas eletrônicas.

Uma função de normalização de valores foi aplicada nos diferentes conjuntos de dados, com objetivo de outorgar o mesmo peso para cada um dos atributos. A normalização padronizou estatisticamente os conjuntos de dados, ajustando os valores dos atributos em um intervalo numérico entre zero e um.

Figura 4 - Ilustração das planilhas eletrônicas resultantes da preparação dos dados.

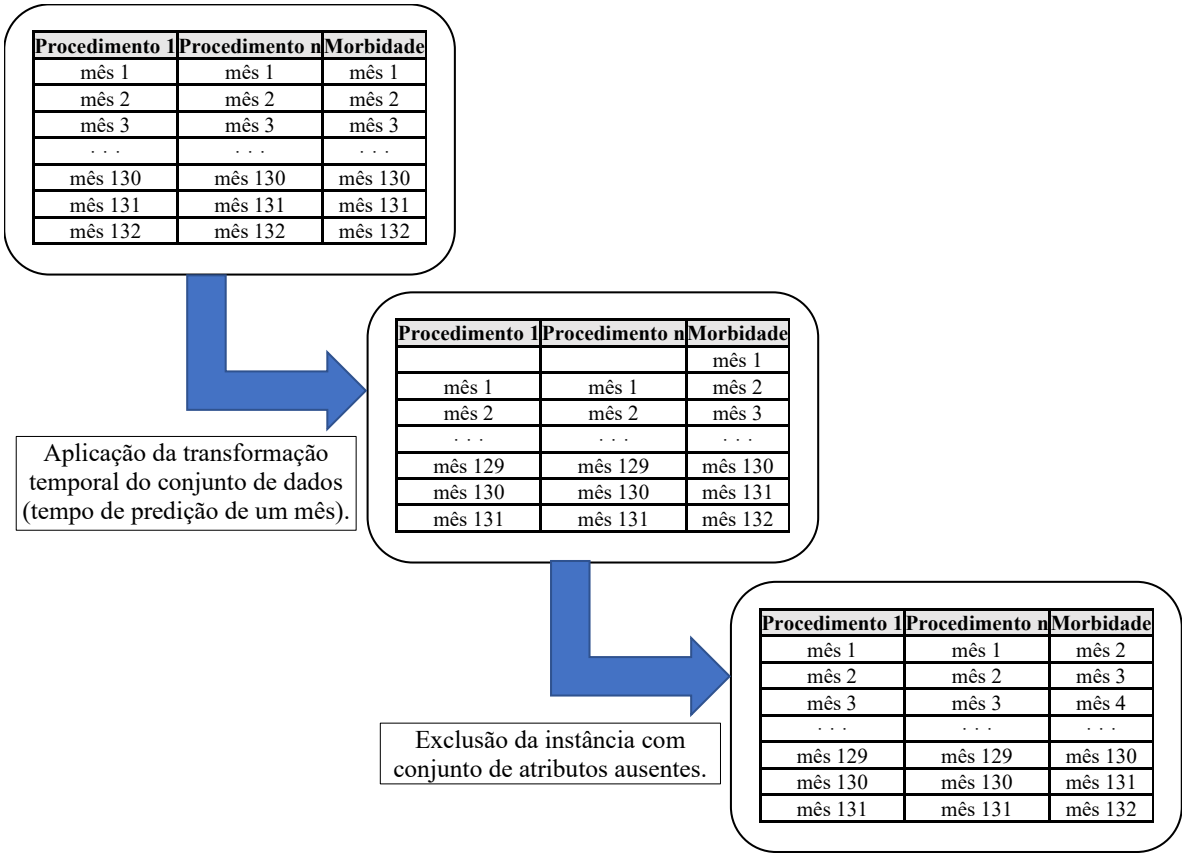


Fonte: Elaborado pelo autor (2019).

A capacidade da predição da morbidade hospitalar foi analisada em diferentes intervalos de tempo: um mês, três meses e seis meses.

Desta forma, em cada conjunto de dados foi aplicada uma transformação temporal que intercalou os atributos com classes de instâncias posteriores, de acordo com cada um dos diferentes períodos analisados. Com a intercalação, uma ou mais instâncias resultaram em conjunto de atributos com valores ausentes, que foram posteriormente excluídas. Consequentemente, os procedimentos ambulatoriais foram instanciados concomitantemente a uma ocorrência futura da morbidade hospitalar, em cada um dos intervalos de predição analisado. Esse processo de preparação dos dados está elucidado na Figura 5. Assim, foram obtidos três diferentes conjuntos de dados para cada um dos 22 capítulos do CID-10, um para cada intervalo de tempo.

Figura 5 - Ilustração dos processos executados na preparação dos dados, para intercalação do período de predição de um mês.



Fonte: Elaborado pelo autor (2019).

Os conjuntos de dados foram exportados para arquivo em formato padrão, denominado *attribute-relation file format* (ARFF), mantendo-os padronizados para a fase de modelagem.

Todos os conjuntos de dados possuíam 490 atributos, que correspondiam a cada um dos procedimentos de diagnóstico em laboratório clínico, mais uma coluna com a classe, ou seja, a morbidade hospitalar. Em razão da relevância heterogênea, repetiu-se a modelagem de dados com a aplicação concomitante de um algoritmo de seleção de atributos. Para tanto, foi executado o algoritmo avaliador de subconjunto de atributos baseado em correlação, com o método de procura Greedy, objetivando a seleção dos atributos mais relevantes para uma melhor modelagem dos dados.

4.3 MODELAGEM

Para a modelagem dos dados foram utilizados os algoritmos:

a) IBk (kNN);
b) M5;
c) Random Forest;
d) SMOReg (SVM); e
e) Multilayer Perceptron.

A modelagem foi executada individualmente com cada algoritmo e, adicionalmente, em conjunto com o processo de seleção de atributos, assim como relatado na preparação dos dados. Os procedimentos foram realizados em cada um dos 66 conjuntos de dados armazenados em arquivos padrão ARFF.

Ainda nessa etapa, o modelo foi avaliado através da aplicação de um processo de validação cruzada de 10-vezes (*10-fold cross-validation*), o qual também foi repetido 10 vezes. Esse procedimento teve o objetivo de garantir resultados mais precisos e minimizar possíveis variações de dados, através das execuções consecutivas. Todos os resultados foram salvos em uma planilha eletrônica padrão CSV.

A métrica utilizada para avaliação dos modelos, através do processo *10-fold cross-validation*, foi o coeficiente de correlação produto-momento ou ρ de Pearson. A métrica foi utilizada para avaliar a correlação entre os valores de predição da quantidade de internações hospitalares, conforme as morbidades agrupadas de acordo com o capítulo do CID-10, e as quantidades reais contidas nos conjuntos de dados separados para teste.

Consequentemente, para testar a hipótese da pesquisa, o coeficiente de correlação foi a variável dependente analisada nos testes estatísticos aplicados, sob a dependência da execução de cada um dos algoritmos e de cada capítulo do CID-10.

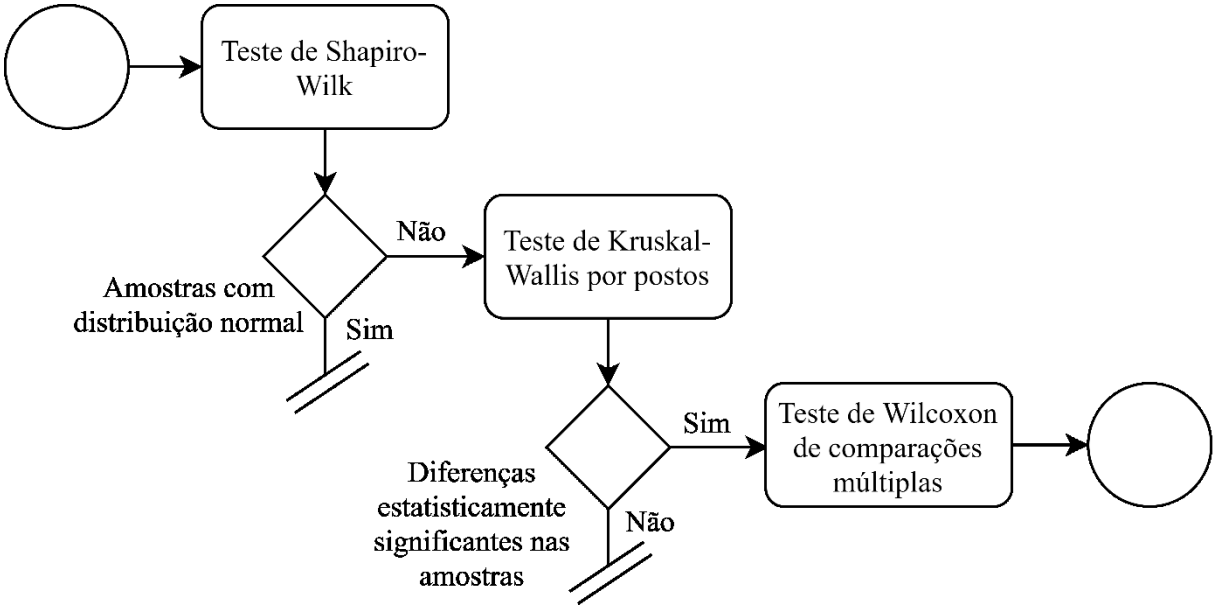
4.4 AVALIAÇÃO

Como objetivo de avaliar os resultados obtidos nos processos de mineração dos dados, foram aplicados testes estatísticos considerando um nível de confiança de 95% (p-valor < 0,05) na variável dependente. Os testes foram realizados em um ambiente de desenvolvimento integrado da linguagem R, aplicando funções de seus pacotes bases.

Os testes estatísticos foram executados separadamente, considerando como amostras individuais os resultados da mineração de dados em cada intervalo de tempo de predição analisado, ou seja, um mês, três meses e seis meses. Os resultados da aplicação dos diferentes testes em cada uma das amostras estão apresentados nos Apêndices.

As amostras continham o coeficiente de correlação ρ de Pearson obtidos nos resultados da execução dos diferentes algoritmos utilizados, associado ou não ao processo de seleção de atributos, nos conjuntos de dados divididos de acordo com os diferentes capítulos do CID-10.

Figura 6 - Diagrama de fluxo dos testes estatísticos executados.



Fonte: Elaborado pelo autor (2019).

Inicialmente foi aplicado um teste estatístico de Shapiro-Wilk para determinar a normalidade das amostras e estabelecer os testes subsequentes. Em seguida, um teste de

Kruskal-Wallis por postos foi utilizado para determinar se existiam diferenças significantes entre os resultados de cada amostra. Por fim, um teste de Wilcoxon de comparações múltiplas foi executado, permitindo a análise dos diferentes algoritmos aplicados em cada um dos capítulos do CID-10 (Figura 6).

O propósito dos testes estatísticos realizados foi determinar quais resultados médios do coeficiente de correlação ρ de Pearson eram diferentes do melhor resultado médio obtido, de acordo com as diferentes comparações analisadas.

5 RESULTADOS

Os resultados dos coeficientes de correção ρ de Pearson, obtidos através dos processos de mineração de dados, estão apresentados na Tabela 1, Tabela 2 e Tabela 3; conforme os intervalos de tempo de predição aplicados na preparação dos dados, respectivamente, um, três e seis meses.

Tabela 1 – Resultados do coeficiente de correlação ρ de Pearson, para predição com intervalo de um mês.

Capítulo CID-10	KNN	KNN + seleção de atributos	M5	M5 + seleção de atributos	Random Forest	Random Forest + seleção de atributos	SVM	SVM + seleção de atributos	Multilayer Perceptron	Multilayer Perceptron + seleção de atributos
I	0,7913	0,7155	0,7438	0,7325	0,7865	**0,8123**	0,6767	0,7556	0,1829	0,7140
II	0,8536	0,8537	0,8439	0,8417	0,8808	0,8821	**0,8882**	0,8866	0,3061	0,8034
III	0,8627	0,8723	0,8337	0,8531	**0,8922**	0,8866	0,8512	0,8863	0,3141	0,7884
IV	0,7675	0,7781	0,7265	0,7568	0,8113	0,7998	0,7664	**0,8127**	0,1387	0,7514
V	0,7891	0,7943	0,8145	0,7894	0,8510	0,8378	**0,8564**	0,8366	0,1337	0,7798
VI	-0,0401	-0,0228	-0,0177	-0,0586	0,0243	0,0189	**0,1432**	-0,0938	0,0795	-0,1033
VII	0,8642	0,8556	0,8528	0,8818	**0,8956**	0,8922	0,8454	0,8954	0,3445	0,8177
VIII	0,2160	0,1156	0,1383	0,1472	**0,3255**	0,2579	0,2557	0,2718	0,0353	0,1928
IX	0,1234	0,2305	0,1726	0,2555	0,2152	0,2987	0,0868	**0,3359**	0,0304	0,2828
X	0,7339	0,6913	0,6972	0,7278	0,6610	**0,7875**	0,4411	0,7165	0,1024	0,7007
XI	0,4969	0,5120	0,5203	0,5585	0,6090	0,6099	0,5975	**0,6285**	0,0916	0,5371
XII	0,8964	0,8827	0,8039	0,8472	0,9159	**0,9164**	0,8639	0,9043	0,3047	0,7636
XIII	0,1479	0,1712	0,1760	0,0985	**0,2578**	0,2024	0,1249	0,2222	-0,0178	0,1824
XIV	0,2775	0,3369	0,3430	0,4523	0,4697	0,4572	0,4237	**0,4828**	0,0970	0,4184
XV	0,3795	0,2573	0,4196	0,2991	0,4205	**0,4339**	0,2418	0,3314	0,0182	0,2465
XVI	0,9548	0,9413	0,9412	0,9261	0,9546	**0,9557***	0,9047	0,9430	0,3687	0,8352
XVII	0,1997	0,1108	0,1474	0,2286	0,2628	**0,2944**	0,2283	0,2707	0,0594	0,1704
XVIII	0,9182	0,9100	0,9031	0,8299	**0,9265**	0,9207	0,8527	0,9168	0,3627	0,8404
XIX	0,9483	0,9418	0,9255	0,9088	**0,9545**	0,9541	0,9268	0,9517	0,3171	0,7807
XX	**0,8962**	0,8479	0,6299	0,8148	0,8266	0,8424	0,5800	0,8056	-0,0856	0,7338
XXI	0,7438	0,7111	0,7568	0,6897	0,7658	**0,8039**	0,4609	0,7416	0,2067	0,6987
XXII	0,4602	0,3068	0,4193	0,4101	**0,4921**	0,4198	0,4787	0,3570	0,1426	0,3177

Fonte: Elaborado pelo autor (2019). Ênfase no maior coeficiente de correlação conforme capítulo do CID-10.
*Melhor resultado do intervalo de tempo.

Os valores médios do coeficiente de correlação ρ de Pearson, obtidos após as repetições do teste de validação cruzada nos modelos produzidos, estão apresentados de acordo

com o capítulo do CID-10 e o algoritmo utilizado. Prospectando uma aplicação prática dos modelos obtidos, os melhores resultados obtidos em cada grupo de doenças (capítulo do CID-10) foram destacados nas tabelas com os resultados individuais.

Tabela 2 - Resultados do coeficiente de correlação ρ de Pearson, para predição com intervalo de três meses.

Capítulo CID-10	KNN	KNN + seleção de atributos	M5	M5 + seleção de atributos	Random Forest	Random Forest + seleção de atributos	SVM	SVM + seleção de atributos	Multilayer Perceptron	Multilayer Perceptron + seleção de atributos
I	**0,7647**	0,6964	0,6421	0,6843	0,7570	0,7523	0,6707	0,7239	0,0654	0,6226
II	0,8602	0,8464	0,7949	0,8031	0,8577	**0,8682**	0,8492	0,8674	0,2022	0,8032
III	0,8714	0,8627	0,8530	0,8443	**0,8817**	0,8774	0,8492	0,8776	0,3052	0,8323
IV	0,7542	0,7616	0,6861	0,7566	0,7956	0,8031	0,7598	**0,8039**	0,0295	0,7156
V	0,7775	0,8008	0,7865	0,7517	0,8353	0,8239	**0,8408**	0,8279	0,0953	0,6662
VI	0,0072	-0,0786	**0,1219**	0,0139	-0,0176	-0,0488	0,1182	0,0102	0,0549	-0,0223
VII	**0,8653**	0,8069	0,8333	0,7461	0,8493	0,8652	0,7836	0,8190	0,3001	0,7676
VIII	0,3201	0,2626	0,3115	0,3074	0,3350	**0,3839**	0,1699	0,3124	0,0651	0,3096
IX	0,3342	0,3113	0,2790	0,3541	0,1881	**0,3918**	0,0888	0,3588	0,0259	0,3386
X	0,7439	0,6809	0,7252	0,7419	0,6982	**0,8292**	0,5508	0,7034	0,1459	0,6631
XI	0,5654	0,5071	0,4771	0,4496	0,5545	**0,5852**	0,5035	0,5451	0,0994	0,5192
XII	0,8991	0,8851	0,8620	0,8588	0,8994	**0,9021**	0,8354	0,8882	0,2530	0,8484
XIII	0,3551	0,3392	0,2642	0,3252	0,2568	**0,4486**	0,0818	0,4183	0,0170	0,4057
XIV	0,3436	0,2810	0,3691	0,2952	**0,4068**	0,3566	0,3351	0,3636	0,0759	0,3033
XV	0,6019	0,5742	0,5988	0,6384	0,7001	**0,7025**	0,4188	0,6461	0,1162	0,5804
XVI	0,9610	0,9560	0,9399	0,9159	0,9623	**0,9682***	0,8952	0,9527	0,3682	0,8815
XVII	0,3850	0,3590	0,4281	**0,4930**	0,3195	0,4849	0,1879	0,4916	0,0465	0,4827
XVIII	**0,9316**	0,9207	0,8911	0,8393	0,9269	0,9306	0,8442	0,9255	0,3989	0,8590
XIX	0,9354	0,9391	0,9287	0,9184	0,9429	0,9417	0,9208	**0,9467**	0,3635	0,9190
XX	**0,8826**	0,8184	0,8483	0,7852	0,8292	0,8170	0,5878	0,7573	-0,1166	0,7257
XXI	**0,7890**	0,6898	0,7125	0,6656	0,7580	0,7484	0,4705	0,7002	0,1582	0,7064
XXII	0,4497	0,4023	0,4862	0,5157	**0,5713**	0,5413	0,5167	0,4459	0,1266	0,4734

Fonte: Elaborado pelo autor (2019). Ênfase no maior coeficiente de correlação conforme capítulo do CID-10. *Melhor resultado do intervalo de tempo.

Para todos os períodos de predição, houve a ocorrência de uma distribuição não normal no conjunto de resultados – tanto em relação aos algoritmos quanto aos capítulos do CID-10 – de acordo com os testes de normalidade Shapiro-Wilk aplicados. Consequentemente, foi demandada a utilização de testes estatísticos que não consideravam como premissa a análise de

dados com distribuição normal. Logo, foram aplicados os testes de Kruskal-Wallis por postos e, complementarmente, os testes de Wilcoxon de comparações múltiplas.

Tabela 3 - Resultados do coeficiente de correlação ρ de Pearson, para predição com intervalo de seis meses.

Capítulo CID-10	KNN	KNN + seleção de atributos	M5	M5 + seleção de atributos	Random Forest	Random Forest + seleção de atributos	SVM	SVM + seleção de atributos	Multilayer Perceptron	Multilayer Perceptron + seleção de atributos
I	0,7753	0,7486	0,6885	0,7357	0,7574	**0,7900**	0,6131	0,7399	0,1143	0,7810
II	0,8495	0,8263	0,8340	0,8215	0,8698	**0,8723**	0,8518	0,8515	0,3279	0,7676
III	0,8653	0,8657	0,8798	0,8623	0,9001	**0,9006**	0,8711	0,8869	0,2070	0,8202
IV	0,7535	0,7627	0,7213	0,7666	0,8027	0,8073	0,7799	**0,8141**	0,0712	0,7517
V	0,7479	0,7818	0,7944	0,7598	0,8175	0,8158	**0,8396**	0,8093	0,1635	0,7221
VI	0,0723	-0,0406	0,0859	0,1131	0,0208	0,0714	0,1128	**0,1993**	-0,0279	0,1115
VII	0,8385	0,7912	0,7197	0,7608	**0,8450**	0,8429	0,7668	0,8279	0,1918	0,7312
VIII	**0,2261**	0,1012	0,1300	0,1456	0,2214	0,1456	0,1513	0,1464	-0,0096	0,0547
IX	0,2699	0,1539	0,2809	0,2056	**0,3188**	0,2427	0,1557	0,1947	-0,0151	0,1873
X	0,6722	0,5489	0,6344	0,6725	0,6352	**0,7208**	0,5206	0,6267	0,0254	0,5462
XI	0,5290	0,4496	0,4843	0,4832	**0,5656**	0,5472	0,5047	0,5483	0,1016	0,4739
XII	0,8916	0,8700	0,8297	0,8310	**0,9031**	0,9018	0,8452	0,8874	0,1668	0,8079
XIII	0,1340	0,0782	0,1248	0,1471	0,1552	0,0835	0,0920	**0,1564**	-0,0280	0,0854
XIV	0,2995	0,1685	0,2625	0,3171	**0,3711**	0,3447	0,3269	0,3454	-0,0169	0,2714
XV	0,4786	0,3977	0,4772	0,4216	0,4401	**0,5011**	0,2158	0,4501	-0,0065	0,3808
XVI	0,9544	0,9416	0,9287	0,9037	**0,9580***	0,9549	0,9143	0,9424	0,2704	0,8806
XVII	0,3060	0,2052	0,3062	0,2136	**0,3121**	0,2253	0,2158	0,2550	0,0279	0,1498
XVIII	0,9261	0,9186	0,9053	0,8843	**0,9340**	0,9306	0,8521	0,9130	0,3484	0,8607
XIX	0,9399	0,9296	0,9160	0,9197	0,9458	**0,9477**	0,9100	0,9379	0,2570	0,8541
XX	**0,8934**	0,8302	0,7752	0,7572	0,8504	0,8368	0,5815	0,8126	-0,0717	0,7687
XXI	0,6744	0,5514	0,5488	0,6401	**0,7194**	0,7021	0,4508	0,6775	0,2364	0,6567
XXII	0,4529	0,3514	0,4479	**0,5204**	0,4850	0,4602	0,5135	0,4791	0,0952	0,5053

Fonte: Elaborado pelo autor (2019). Ênfase no maior coeficiente de correlação conforme capítulo do CID-10. *Melhor resultado do intervalo de tempo.

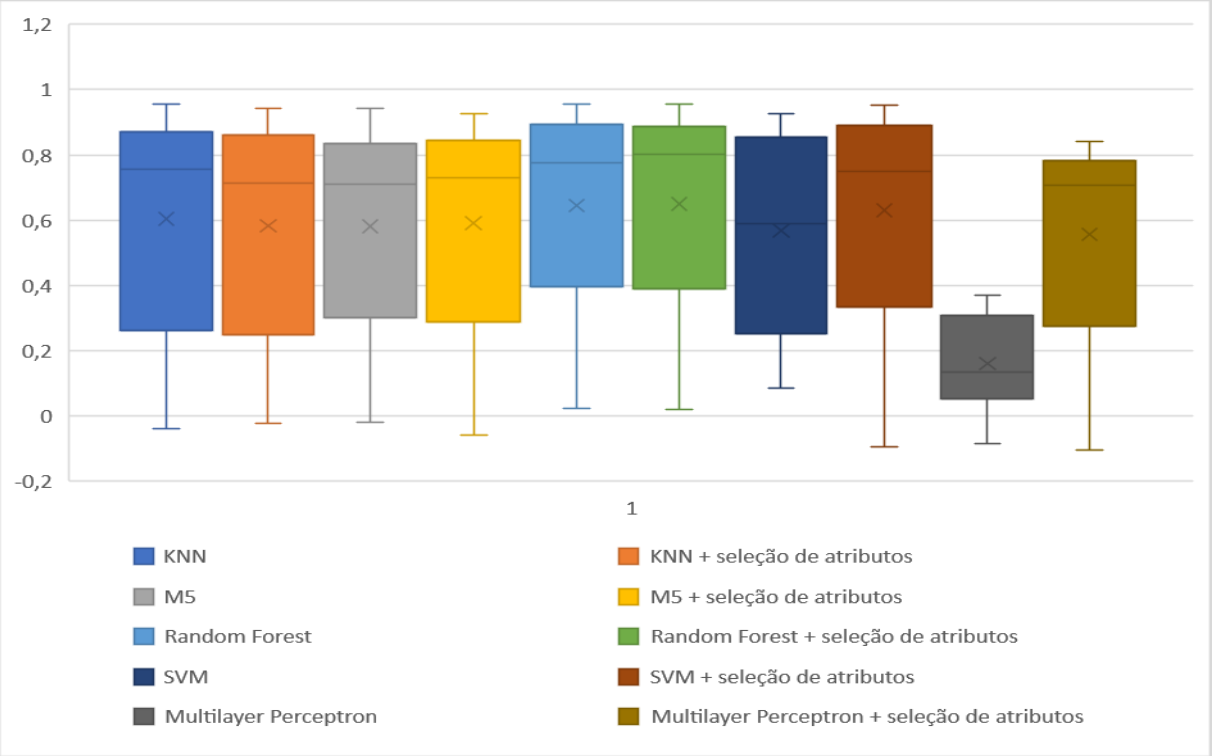
Os testes de Kruskal-Wallis nos grupos de resultados – tanto em relação ao algoritmo utilizado quanto ao capítulo do CID-10 considerado – resultaram em diferenças estatisticamente significante em cada um dos períodos de predição analisado. Desta forma, os testes de Wilcoxon de comparações múltiplas foram necessários para identificar quais resultados possuíam diferenças significantes.

As análises estatísticas de comparações múltiplas permitiram determinar os resultados que possuíam diferenças significantes, entre aqueles melhores na amostra das quais faziam parte. Assim, foi possível realizar uma análise individual das relações entre as variáveis estudadas nessa pesquisa, de acordo com cada intervalo de tempo de predição.

Considerando todos os intervalos de predição, os algoritmos que obtiveram os melhores resultados, tomando como perspectiva os capítulos do CID-10, foram: M5 (em apenas um conjunto de dados), M5 associado à seleção de atributos (em dois conjuntos de dados), SVM (em cinco conjuntos de dados), kNN (em oito conjuntos de dados), SVM associado à seleção de atributos (em nove conjuntos de dados), Random Forest (em 19 conjuntos de dados) e Random Forest associado à seleção de atributos (em 22 conjuntos de dados).

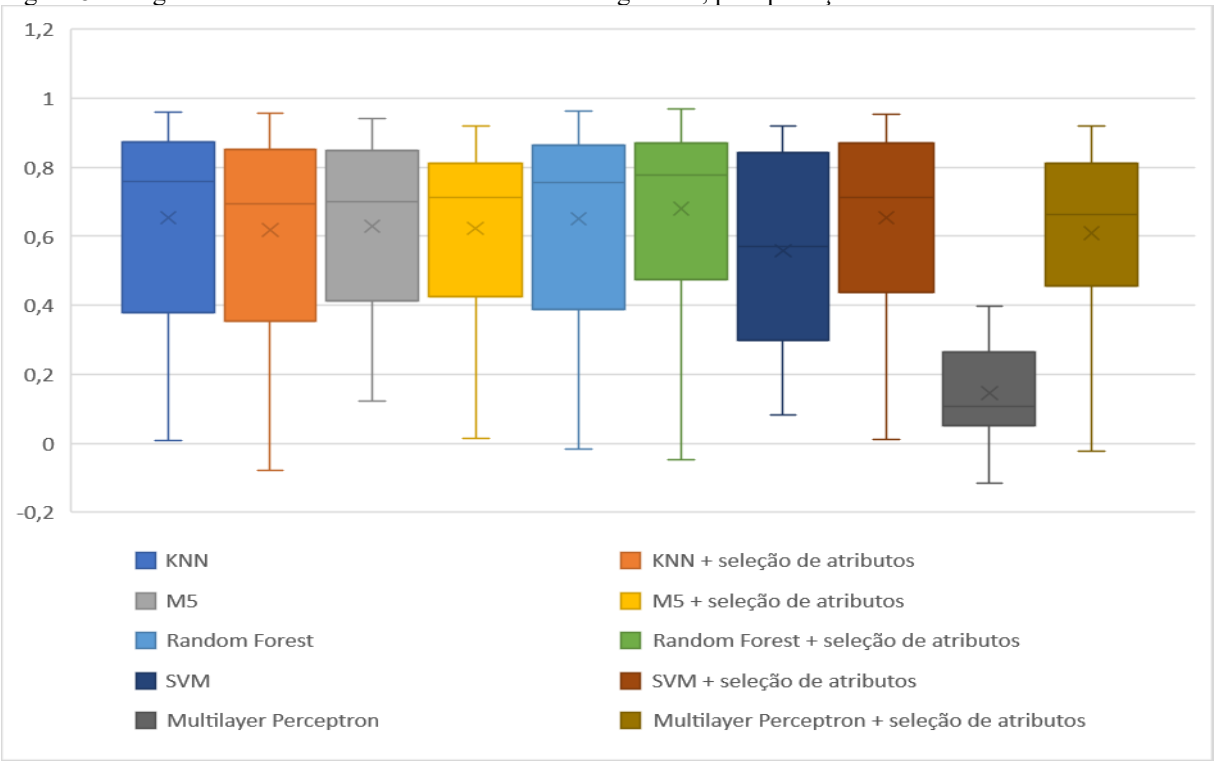
A variação entre os grupos de resultados, ou seja, nos diferentes intervalos de tempo de predição, está expressa nos diagramas de caixas (*bloxpots*) ilustrados na Figura 7, Figura 8 e Figura 9, conforme o algoritmo utilizado para modelagem dos dados.

Figura 7 - Diagrama de caixas dos resultados conforme algoritmo, para predição com intervalo de um mês.



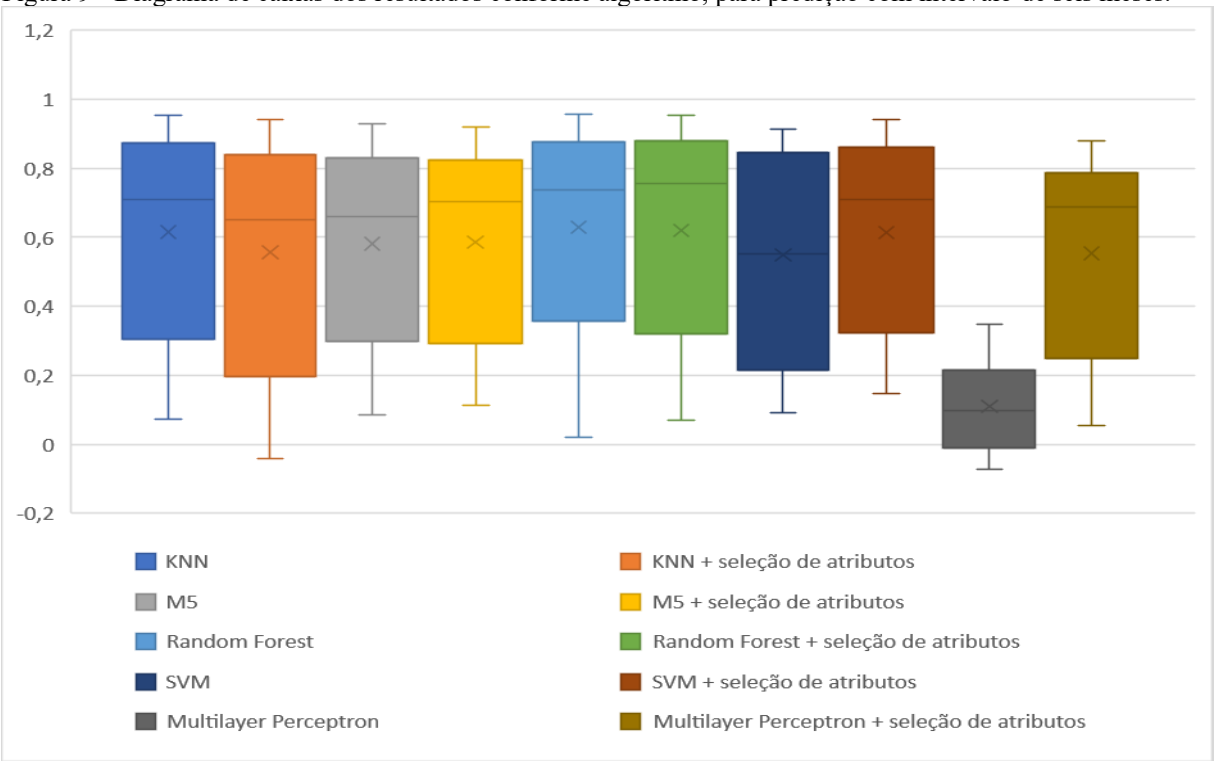
Fonte: Elaborado pelo autor (2019).

Figura 8 - Diagrama de caixas dos resultados conforme algoritmo, para predição com intervalo de três meses.



Fonte: Elaborado pelo autor (2019).

Figura 9 - Diagrama de caixas dos resultados conforme algoritmo, para predição com intervalo de seis meses.



Fonte: Elaborado pelo autor (2019).

Os *bloxpots* apresentados permitem uma melhor comparação da variação dos resultados obtidos, principalmente na presença dos valores atípicos existentes em todos os intervalos de

tempo, bem como os quartis dos conjuntos de resultados conforme algoritmo empregado. Desta forma, a seguir estão as principais informações acerca dos resultados obtidos, de acordo com o intervalo de tempo de predição, sob a perspectiva dos algoritmos utilizados e os capítulos do CID-10.

5.1 RESULTADOS PARA PREDIÇÃO COM INTERVALO DE UM MÊS

Para a predição com intervalo de tempo de um mês, o resultado médio do coeficiente de correlação ρ de Pearson, obtido pela repetição dos testes de validação, foi de 0,5568. Esse resultado pode ser categorizado como uma correlação moderada, de acordo com a classificação proposta por Evans (1996).

O algoritmo que apresentou o melhor resultado médio foi Random Forest associado com a seleção de atributos, com o coeficiente de correlação de 0,6493, categorizado como uma correlação moderada. De acordo com o teste de comparações múltiplas, o resultado médio da utilização do algoritmo Random Forest com a seleção de atributos não possuía diferenças significantes entre todos os outros, com exceção da aplicação do Multilayer Perceptron sem o uso do algoritmo de seleção de atributos. A categorização dos resultados médios da aplicação de cada algoritmo pode ser averiguada na Tabela 4.

Tabela 4 - Resultados médios do coeficiente de correlação ρ de Pearson, por algoritmo (intervalo de um mês).

Algoritmo	**Resultado médio**	**Categorização conforme Evans**
Random Forest + seleção de atributos	0,6493	Correlação moderada
Random Forest	0,6454	Correlação moderada
SVM + seleção de atributos	0,6300	Correlação moderada
KNN	0,6037	Correlação moderada
M5 + seleção de atributos	0,5905	Correlação moderada
KNN + seleção de atributos	0,5824	Correlação moderada
M5	0,5814	Correlação moderada
SVM	0,5680	Correlação moderada
Multilayer Perceptron + seleção de atributos	0,5569	Correlação moderada
Multilayer Perceptron	0,1606*	Correlação muito fraca

Fonte: Elaborado pelo autor (2019). *Diferença estatisticamente significante em relação ao melhor resultado.

Tabela 5 - Resultados médios do coeficiente de correlação ρ de Pearson, por capítulo do CID-10 (intervalo de um mês).

Capítulo do CID-10	Resultado médio	Categorização conforme Evans
Algumas afecções originadas no período perinatal	0,8725	Correlação muito forte
Lesões, envenenamentos e algumas outras consequências de causas externas	0,8609	Correlação muito forte
Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	0,8381	Correlação muito forte
Doenças do olho e anexos	0,8145	Correlação muito forte
Doenças da pele e do tecido subcutâneo	0,8099	Correlação muito forte
Neoplasmas (tumores)	0,8040	Correlação muito forte
Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários	0,8040	Correlação muito forte
Transtornos mentais e comportamentais	0,7483	Correlação forte
Doenças endócrinas, nutricionais e metabólicas	0,7109	Correlação forte
Algumas doenças infecciosas e parasitárias	0,6911	Correlação forte
Causas externas de morbidade e de mortalidade	0,6892	Correlação forte
Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	0,6579	Correlação forte
Doenças do aparelho respiratório	0,6259	Correlação forte
Doenças do aparelho digestivo	0,5161	Correlação moderada
Códigos para propósitos especiais	0,3804	Correlação fraca
Doenças do aparelho geniturinário	0,3759	Correlação fraca
Gravidez, parto e puerpério	0,3048*	Correlação fraca
Doenças do aparelho circulatório	0,2032*	Correlação fraca
Malformações congênitas, deformidades e anomalias cromossômicas	0,1973*	Correlação muito fraca
Doenças do ouvido e da apófise mastoide	0,1956*	Correlação muito fraca
Doenças do sistema osteomuscular e do tecido conjuntivo	0,1565*	Correlação muito fraca
Doenças do sistema nervoso	-0,0070*	Correlação muito fraca

Fonte: Elaborado pelo autor (2019). *Diferença estatisticamente significante em relação ao melhor resultado.

Em relação ao CID-10, o capítulo XVI (Algumas afecções originadas no período perinatal) apresentou a melhor média do coeficiente correlação, com resultado de 0,8725, podendo ser categorizado como uma correlação muito forte. Os capítulos VI (Doenças do sistema nervoso), VIII (Doenças do ouvido e da apófise mastoide), IX (Doenças do aparelho circulatório), XIII (Doenças do sistema osteomuscular e do tecido conjuntivo), XV (Gravidez,

parto e puerpério) e XVII (Malformações congênitas, deformidades e anomalias cromossômicas) possuíam diferenças significantes, ou seja, apresentaram resultados de correlação menores.

A classificação dos resultados médios obtidos, conforme os capítulos do CID-10, está demonstrada na Tabela 5.

O maior coeficiente de correlação alcançado foi a aplicação do algoritmo Random Forest associado a seleção de atributos na predição de internações do capítulo XVI (Algumas afecções originadas no período perinatal), com resultado de 0,9557, classificado como correlação muito forte. Os resultados individuais da aplicação de cada algoritmo, em todas as classificações do CID-10, podem também ser visualizados na Tabela 1.

5.2 RESULTADOS PARA PREDIÇÃO COM INTERVALO DE TRÊS MESES

As repetições dos testes nos modelos produzidos resultaram em uma média do coeficiente de correlação ρ de Pearson de 0,5823, para o intervalo de tempo de três meses. Conforme Evans (1996) sugeriu, o resultado pode ser considerado como correlação moderada.

Tabela 6 - Resultados médios do coeficiente de correlação ρ de Pearson, por algoritmo (intervalo de três meses).

Algoritmo	Resultado médio	Categorização conforme Evans
Random Forest + seleção de atributos	0,6806	Correlação moderada
KNN	0,6545	Correlação moderada
SVM + seleção de atributos	0,6539	Correlação moderada
Random Forest	0,6504	Correlação moderada
M5	0,6291	Correlação moderada
M5 + seleção de atributos	0,6229	Correlação moderada
KNN + seleção de atributos	0,6192	Correlação moderada
Multilayer Perceptron + seleção de atributos	0,6091	Correlação moderada
SVM	0,5581	Correlação moderada
Multilayer Perceptron	0,1453*	Correlação muito fraca

Fonte: Elaborado pelo autor (2019). *Diferença estatisticamente significante em relação ao melhor resultado.

Os resultados médios dos algoritmos são apresentados na Tabela 6. O algoritmo que apresentou o melhor resultado médio foi Random Forest em conjunto com a seleção de atributos, com o coeficiente de correlação de 0,6806. Este resultado pode ser considerado como

uma correlação moderada, diferente estatisticamente apenas do algoritmo Multilayer Perceptron.

Dos resultados médios, quando considerado cada capítulo do CID-10, foi possível verificar uma correlação de 0,8801 no capítulo XVI (Algumas afecções originadas no período perinatal) como o maior resultado. Considerado uma correlação forte, esse resultado médio possuía diferença estatisticamente significante entre: Doenças do sistema nervoso (capítulo VI); Doenças do ouvido e da apófise mastoide (capítulo VIII); Doenças do aparelho circulatório (capítulo IX); Doenças do sistema osteomuscular e do tecido conjuntivo (capítulo XIII); Doenças do aparelho geniturinário (capítulo XIV); e Malformações congênitas, deformidades e anomalias cromossômicas (capítulo XVII).

Em relação à classificação quanto ao capítulo do CID-10, os resultados estão dispostos na Tabela 7.

Tabela 7 - Resultados médios do coeficiente de correlação ρ de Pearson, por capítulo do CID-10 (intervalo de três meses).

Capítulo do CID-10	Resultado médio	Categorização conforme Evans
Algumas afecções originadas no período perinatal	0,8801	Correlação muito forte
Lesões, envenenamentos e algumas outras consequências de causas externas	0,8756	Correlação muito forte
Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	0,8468	Correlação muito forte
Doenças da pele e do tecido subcutâneo	0,8132	Correlação muito forte
Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários	0,8055	Correlação muito forte
Neoplasmas (tumores)	0,7752	Correlação forte
Doenças do olho e anexos	0,7636	Correlação forte
Transtornos mentais e comportamentais	0,7206	Correlação forte
Causas externas de morbidade e de mortalidade	0,6935	Correlação forte
Doenças endócrinas, nutricionais e metabólicas	0,6866	Correlação forte
Doenças do aparelho respiratório	0,6483	Correlação forte
Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	0,6399	Correlação forte
Algumas doenças infecciosas e parasitárias	0,6379	Correlação forte
Gravidez, parto e puerpério	0,5577	Correlação moderada
Doenças do aparelho digestivo	0,4806	Correlação moderada
Códigos para propósitos especiais	0,4529	Correlação moderada

Capítulo do CID-10	Resultado médio	Categorização conforme Evans
Malformações congênitas, deformidades e anomalias cromossômicas	0,3678*	Correlação fraca
Doenças do aparelho geniturinário	0,3130*	Correlação fraca
Doenças do sistema osteomuscular e do tecido conjuntivo	0,2912*	Correlação fraca
Doenças do ouvido e da apófise mastoide	0,2778*	Correlação fraca
Doenças do aparelho circulatório	0,2671*	Correlação fraca
Doenças do sistema nervoso	0,0159*	Correlação muito fraca

Fonte: Elaborado pelo autor (2019). *Diferença estatisticamente significante em relação ao melhor resultado.

Dos resultados individuais, o maior coeficiente de correlação foi obtido com a aplicação do algoritmo Random Forest associado à seleção de atributos. No capítulo XVI (Algumas afecções originadas no período perinatal) a correlação foi de 0,9682. Este foi o melhor resultado individual obtido entre todos os testes realizados, considerado uma correlação muito forte. Os resultados individuais podem ser averiguados na Tabela 2.

5.3 RESULTADOS PARA PREDIÇÃO COM INTERVALO DE SEIS MESES

Para a predição com intervalo de tempo de seis meses, a média do coeficiente de correlação ρ de Pearson foi de 0,5414, que pode ser considerada uma correlação moderada (EVANS, 1996).

Resultado classificado como correlação moderada, o algoritmo Random Forest apresentou o melhor valor médio, com o coeficiente de correlação 0,6286. O resultado da aplicação desse algoritmo não possuía diferenças significantes entre todos os outros, com exceção da aplicação de Multilayer Perceptron sem o uso do algoritmo de seleção de atributos. Os valores médios dos coeficientes de correlação da aplicação de cada algoritmo podem ser averiguados na Tabela 8.

Tabela 8 - Resultados médios do coeficiente de correlação ρ de Pearson, por algoritmo (intervalo de seis meses).

Algoritmo	Resultado médio	Categorização conforme Evans
Random Forest	0,6286	Correlação moderada
Random Forest + seleção de atributos	0,6202	Correlação moderada
KNN	0,6159	Correlação moderada

Algoritmo	Resultado médio	Categorização conforme Evans
SVM + seleção de atributos	0,6137	Correlação moderada
M5 + seleção de atributos	0,5856	Correlação moderada
M5	0,5807	Correlação moderada
KNN + seleção de atributos	0,5560	Correlação moderada
Multilayer Perceptron + seleção de atributos	0,5531	Correlação moderada
SVM	0,5493	Correlação moderada
Multilayer Perceptron	0,1104*	Correlação muito fraca

Fonte: Elaborado pelo autor (2019). *Diferença estatisticamente significante em relação ao melhor resultado.

O capítulo XVI (Algumas afecções originadas no período perinatal) apresentou a melhor média do coeficiente correlação, com resultado de 0,8649, classificado como correlação forte. Os capítulos que possuíam diferenças significantes eram: capítulos VI (Doenças do sistema nervoso), VIII (Doenças do ouvido e da apófise mastoide), IX (Doenças do aparelho circulatório), XIII (Doenças do sistema osteomuscular e do tecido conjuntivo) e XVII (Malformações congênitas, deformidades e anomalias cromossômicas). Entre esses, a média de correlação foi menor do que o capítulo XVI.

A classificação dos resultados médios obtidos, conforme os capítulos do CID-10, está demonstrada na Tabela 9.

Tabela 9 - Resultados médios do coeficiente de correlação ρ de Pearson, por capítulo do CID-10 (intervalo de seis meses).

Capítulo do CID-10	Resultado médio	Categorização conforme Evans
Algumas afecções originadas no período perinatal	0,8649	Correlação muito forte
Lesões, envenenamentos e algumas outras consequências de causas externas	0,8558	Correlação muito forte
Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	0,8473	Correlação muito forte
Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários	0,8059	Correlação muito forte
Doenças da pele e do tecido subcutâneo	0,7935	Correlação forte
Neoplasmas (tumores)	0,7872	Correlação forte
Doenças do olho e anexos	0,7316	Correlação forte
Transtornos mentais e comportamentais	0,7252	Correlação forte
Causas externas de morbidade e de mortalidade	0,7034	Correlação forte

Capítulo do CID-10	Resultado médio	Categorização conforme Evans
Doenças endócrinas, nutricionais e metabólicas	0,7031	Correlação forte
Algumas doenças infecciosas e parasitárias	0,6744	Correlação forte
Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	0,5858	Correlação moderada
Doenças do aparelho respiratório	0,5603	Correlação moderada
Doenças do aparelho digestivo	0,4688	Correlação moderada
Códigos para propósitos especiais	0,4311	Correlação moderada
Gravidez, parto e puerpério	0,3756	Correlação fraca
Doenças do aparelho geniturinário	0,2690*	Correlação fraca
Malformações congênitas, deformidades e anomalias cromossômicas	0,2217*	Correlação fraca
Doenças do aparelho circulatório	0,1994*	Correlação muito fraca
Doenças do ouvido e da apófise mastoide	0,1313*	Correlação muito fraca
Doenças do sistema osteomuscular e do tecido conjuntivo	0,1029*	Correlação muito fraca
Doenças do sistema nervoso	0,0719*	Correlação muito fraca

Fonte: Elaborado pelo autor (2019). *Diferença estatisticamente significante em relação ao melhor resultado.

O melhor coeficiente de correlação nesse intervalo de tempo foi obtido através da aplicação do algoritmo Random Forest no capítulo XVI (Algumas afecções originadas no período perinatal), com resultado de 0,9580 e classificado como correlação muito forte. Os resultados individuais podem novamente serem constatados na Tabela 3.

6 DISCUSSÃO

A principal premissa dessa pesquisa é alicerçada na consagrada história natural da doença proposta por Leavell e Clark (1976), que ainda é afirmativa extensivamente utilizada em Ciências da Saúde, principalmente em Epidemiologia. Doenças evoluem: do período de pré-patogênese para o período de patogênese; sincronicamente, sob outra perspectiva, do nível de prevenção primária para a prevenção terciária.

Consequentemente, considerando os aspectos dessa pesquisa e ainda utilizando os mesmos conceitos básicos, pacientes realizam exames seletivos para o diagnóstico precoce, usualmente em ambientes ambulatoriais. Não havendo medidas que alterem a história da doença, evoluem para estágios mais avançados e para a convalescença, demandando atenção à saúde de maior complexidade, mormente serviços hospitalares.

Logo, se existe um percurso na evolução de uma doença, seria possível a predição da morbidade hospitalar considerando a informação disponível do atendimento ambulatorial, sob uma perspectiva epidemiológica? Partindo dessa pergunta, foi formulada a hipótese de pesquisa sobre a possibilidade de construção de modelos que permitiriam a predição de morbidade hospitalar através dos dados de registro de atendimento ambulatorial, em razão da existência de uma possível correlação até então desconhecida.

Outra circunstância considerada importante, os dados utilizados eram provenientes de fontes secundárias, amplamente disseminados e de apresentação obrigatória no sistema de saúde pública brasileiro. Da mesma forma, esses continham informações de todo o país e uma série histórica de ao menos onze anos, bem como estavam armazenados em um mesmo repositório, sem a identificação pessoal de pacientes e facilmente acessíveis (BRASIL, 2010).

Miotto et al. (2016) propuseram a utilização de resultados de exames laboratoriais para melhorar a capacidade preditiva de um modelo de dados pesquisados por eles, construído a partir dos registros eletrônicos de saúde, com o objetivo de prever doenças em pacientes. Porém, ressaltaram que existiam dificuldades no processamento desse tipo de informação. Os principais problemas estariam relacionados à integridade dos dados utilizados, referentes aos resultados dos procedimentos de diagnóstico em laboratório clínico; como também, em razão da possível ausência de confidencialidade e dificuldades em relação à sua disponibilidade.

Outrossim, para que a mineração de dados tenha capacidade de melhorar a precisão do diagnóstico e possuir capacidade preditiva para suporte à decisão, sendo aceito nos processos de trabalho de saúde, os bancos de dados devem ser padronizados e as informações precisam

ser passíveis de compartilhamento. Adicionalmente a essas características, esses mesmos dados de saúde devem ser coletados prospectivamente (GILLIES et al., 2015).

Consequentemente, os dados utilizados na presente pesquisa obrigatoriamente precisariam deter características importantes relacionadas à segurança da informação: confidencialidade, integridade e disponibilidade. Esse aspecto foi determinante para a execução dos procedimentos e para os propósitos do futuro desenvolvimento de uma aplicação prática. Sob uma perspectiva de implementação dos modelos obtidos em sistemas de informação, essas propriedades permitem uma aplicação irrestrita e, principalmente, sustentável.

Conquanto nem todas outras informações contidas nos bancos de dados do DATASUS estarem suficientemente estruturadas para a geração de conhecimento, os dados utilizados para a construção dos modelos eram provenientes do subgrupo de procedimentos de diagnóstico em laboratório clínico. Esses dados são os mais estruturados e detalhados no sistema de informação SIASUS: o subgrupo possui 490 procedimentos, organizados em 12 subconjuntos diferentes (formas de organização) (BRASIL, 2007).

Igualmente, a morbidade relacionada à internação hospitalar é informação obrigatória nos registros do SIHSUS. A Classificação Internacional de Doenças é padrão internacionalmente adotado e instituído pela Organização Mundial de Saúde (WORLD HEALTH ORGANIZATION, 2019). Bem como, diversos autores empregam os códigos CID-10 para determinar qual a condição de saúde que um paciente carrega, quando desenvolvem suas pesquisas (FERLAY et al., 2018; RICHTER; KHOSHGOFTAAR, 2018; SANTOS; CARVALHO, 2018; SHICKEL et al., 2018).

Justamente pela existência dessas características, foi aventada a hipótese da construção de modelos para a geração de conhecimento útil, através da aplicação de algoritmos de aprendizagem de máquina especificamente nesses dados. Desta forma seria possível alcançar um maior ganho de informação nos processos de mineração de dados e melhor entendimento dos resultados obtidos.

6.1 MINERAÇÃO DE DADOS EM SAÚDE

Apesar de se tratar de dados temporais, essa pesquisa não utilizou métodos de modelagem que considerassem a sucessão de dados no tempo. A aplicação de processos específicos para informação vinculadas ao tempo acrescentaria mais variáveis na análise da correlação entre os modelos de dados ambulatoriais e os dados hospitalares futuros. Por exemplo, certos grupos de doenças possuem alguma relação cíclica anual, principalmente

relacionados à estação climática do ano, como é o caso de algumas doenças infecciosas. De outra maneira, doenças relacionadas a eventos sem relação cíclica, como as afecções originadas no período perinatal, não sofreriam essa influência.

Ora, dados cíclicos também podem ser avaliados através de métodos estatísticos de regressão, sem a necessidade da utilização da aprendizagem de máquina (WITTEN et al., 2016; SANTOS; CARVALHO, 2018). Como o propósito dos testes realizados era a analisar a correlação entre um modelo de dados ambulatoriais construído e dados de morbidade hospitalar, a execução de procedimentos que considerassem a relação temporal dos dados poderia interferir nessa finalidade. Todavia, diversos estudos obtiveram resultados importantes na mineração de dados aplicando algoritmos que modelam a relação temporal (CUI et al., 2016; ALSHAHRANI et al., 2018; SHICKEL et al., 2018; FAWAZ et al., 2019; LIU et al., 2019). E essas técnicas devem ser objeto de pesquisas futuras nesses mesmos dados, principalmente quando forem aplicadas em um subgrupo ou doença específica.

Sob outra perspectiva, algoritmos de aprendizagem de máquina podem ser considerados transparentes ou não - no último caso, chamados de "*black boxes*", como Random Forest, SVM e as redes neurais artificiais utilizadas nessa pesquisa. Autores ressaltam que os algoritmos *black boxes* são mais robustos e possibilitam a obtenção de resultados notadamente melhores do que os outros, como exemplo, as redes neurais artificiais que aplicam *deep learning* (BIBAULT et al., 2016; MIOTTO et al., 2016; RAVÌ et al., 2017; SHICKEL et al., 2018; FAWAZ et al., 2019). Contudo, assim como ponderado por Breiman (2001), El Naqa (2016), Witten et al. (2016), Afzal et al. (2017), Belinger et al. (2017), Camacho (2018), Ching et al. (2018), Mlakar et al. (2018) e Shickel et al. (2018), nas diferentes aplicações de mineração de dados em suas pesquisas, interpretar os processos que conduziram ao conhecimento gerado deve também ser possível, para que a utilização de modelos de dados na tomada de decisão no diagnóstico de doenças e estabelecimento de tratamento para pacientes seja amplamente utilizado na clínica diária.

Modelos devem ser transparentes para que os profissionais de saúde tenham informação suficiente para alicerçar a correta tomada de decisão. Mlakar et al. (2018) pondera que algoritmos "*white boxes*" (transparentes) são melhor empregados como ferramenta especializada e analítica de suporte à decisão; elucidando a importância do entendimento dos procedimentos envolvidos na mineração de dados, ao permitir sua aplicabilidade efetiva nos diferentes campos de Ciência da Saúde. Entre os algoritmos utilizados na presente pesquisa, são considerados transparentes o M5 e o kNN.

A interpretabilidade dos modelos de dados é importante para proporcionar confiança para o diagnóstico em saúde, permitindo decisões por motivos conhecidos e não por um artefato dos dados. Modelos preditivos que geram conhecimento inovador devem permitir a identificação dos padrões de dados que chegaram ao resultado, com o propósito de possibilitar sua compreensão pelos profissionais da sua área de aplicação (CHING et al., 2018). Afinal, processos de modelagem de dados para evidenciação de informação podem ser executados por diferentes algoritmos, aplicando os conceitos de Inteligência Artificial; todavia, até o presente momento, conhecimento ainda precisa ser gerado pela humanidade, mesmo porque existem aspectos legais envolvidos nessas situações.

Para a resolução diagnóstica e a definição do tratamento com a melhor relação entre vantagens e desvantagens – ou custo-benefício – devem ser considerados diversos fatores que extrapolam apenas a análise das informações disponíveis. Profissionais de saúde remetem à sua experiência – e a de seus pares – e às demandas de seus pacientes para formar opinião sobre um caso clínico e realizar a tomada de decisão. Consequentemente, seria raso sugerir que para a utilização de um modelo obtido por aprendizagem de máquina sejam considerados apenas os seus resultados finais, mesmo que sejam excelentes, e não a plena interpretação dos processos e cálculos que nele resultaram. Esse paradigma deve nortear pesquisas futuras para obtenção de modelos que atendam aos problemas que sua construção propõe, ou, de uma outra forma, que sejam efetivamente utilizados para a tomada de decisão em saúde.

O estado-da-arte em mineração de dados exige o constante desenvolvimento de técnicas e algoritmos, que acompanham a evolução da capacidade de processamento de computadores e o acúmulo de dados disponíveis. Esse constante desenvolvimento também é exigido dos profissionais, nas diferentes áreas do conhecimento, para a extração da maior quantidade de conhecimento possível de um conjunto de dados. Contudo, Chan et al. (2017) ressalta que os resultados da mineração de dados são alicerçados pela interpretação das informações geradas. Em outras palavras, conforme os autores, o desafio é específico do domínio e não técnico.

Essa conjuntura explica a importância da compreensão da mineração de dados por profissionais de saúde, principalmente porque a interpretação das informações e a efetiva utilização do conhecimento gerado depende da participação ativa dos profissionais da área de conhecimento dos dados analisados. Profissionais de Ciência de Computação são indutores importantes para o desenvolvimento dos algoritmos e ferramentas, assim como implementação de sistemas inteligentes; contudo, não lhes cabe a interpretação do conhecimento quando esse se refere a uma diferente área do conhecimento. E, talvez, o próximo evento disruptivo seja a

proficiência da mineração de dados, seu ensino na formação profissional e acadêmica, em todas as áreas do conhecimento.

6.2 INTERPRETAÇÃO DOS RESULTADOS

A pesquisa teve o propósito de tratar um problema de regressão. Zhao (2012) sintetiza que a regressão constrói uma função de atributos independentes (frequência de procedimentos ambulatoriais de diagnóstico em laboratório clínico) para prever um atributo dependente (frequência de morbidades registradas, por capítulo do CID-10, em atendimentos hospitalares). É importante estabelecer um discernimento entre esses atributos do conjunto de dados e as variáveis que são objetos de análise dessa pesquisa: os algoritmos utilizados na modelagem e os capítulos do CID-10 – variáveis independentes –, assim como o coeficiente de correlação ρ de Pearson – variável dependente.

Diversos autores relatam excelentes resultados de mineração de dados com a utilização de algoritmos como Random Forest, que executam processos de modelagem baseados em *ensemble learning* (MIOTTO et al., 2016; ALTHUNAYAN et al., 2017; AWAD et al. 2017; KUMAR; KHATRI, 2017; BAŞAR; AKAN, 2018; BORAH et al., 2018; FARZI et al., 2018; JHA et al., 2018; MLAKAR et al., 2018; SAHNI et al., 2018; SHIMODA et al., 2018; KAUR et al., 2019). Porém, usualmente são esperados melhores resultados com algoritmos mais sofisticados, como é o caso das redes neurais artificiais (RICHTER; KHOSHGOFTAAR, 2018). Contrariando os resultados da presente pesquisa, em que o algoritmo Multilayer Perceptron apresentou os piores resultados (com diferenças estatisticamente significantes sem a aplicação do algoritmo de seleção de atributos).

Witten et al. (2016) ressaltaram alguns problemas das redes neurais artificiais, que podem explicar o mal resultado obtido por esse algoritmo. Entre esses, se a função de descida do gradiente calcula vários pontos de erros mínimos locais, o algoritmo encontra dificuldades para identificar o erro mínimo global e, consequentemente, obter modelos de dados adequados para a predição. Outrossim, algoritmos de aprendizagem de máquina mais complexo, como o caso das redes neurais artificiais, são "*data hungry*", ou seja, necessitam de grande quantidade de instâncias de dados para alcançar o nível adequado de treinamento e construir modelos de dados adequados (VAN DER PLOEG et al. 2014; CAMACHO et al., 2018)

Da mesma forma, poderia ser esperado que o algoritmo SVM alcançasse melhores resultados do que o Random Forest (AWAD et al., 2017; SRIVASTAVA et al., 2017; SANEJA; RANI, 2019). De fato, as máquinas de vetores de suporte possuem excelente capacidade de

evitar o sobre-ajuste – em inglês, *overfitting* –, ou seja, quando um modelo se ajusta tão bem ao conjunto de dados observado que se torna inadequado para realizar predição em novas instâncias. Todavia, a alteração de poucos vetores de suporte pode ocasionar oscilações marcantes em grandes seções do limite de decisão, comprometendo os resultados do algoritmo (WITTEN et al., 2016). Ainda assim, a utilização em conjunto com o algoritmo de seleção de atributos permitiu uma melhora significativa na performance do algoritmo SVM em todos os intervalos de tempo de predição.

A característica de alta dimensionalidade do conjunto de dados pode ter influenciado os resultados apresentados nessa pesquisa. Sob a perspectiva da quantidade de atributos, a performance de determinados algoritmos pode ficar comprometida quando são aplicados em um conjunto de dados com grande dimensionalidade. Na pesquisa o conjunto de dados possuía 490 atributos mais a classe.

Por outro lado, alguns algoritmos intrinsicamente incorporam processos de seleção de atributos, como é o caso das árvores de decisão (BREINMAN, 2001; HUSSIEN et al., 2017). Em um conjunto de dados com muitos atributos, o processo pode ter conduzido a uma seleção daqueles que possuem maior relação com a classe. Nos processos realizados nessa pesquisa, o algoritmo intrinsicamente selecionou os procedimentos ambulatoriais de diagnóstico em laboratório clínico que mais afetavam a frequência de morbidades hospitalares registradas por capítulo do CID-10. Essa afirmativa também alicerça a explicação dos melhores resultados alcançados com a aplicação do algoritmo de seleção de atributos associado à modelagem dos dados.

Random Forest já demonstrou ser adequado para modelagem de dados com grande dimensionalidade. Além de ser menos sensível à qualidade de amostras de treinamento e ao *overfitting*, quando comparado a outros algoritmos. Isso devido ao grande número de árvores de decisão produzidas durante a seleção aleatória do subconjunto de amostras de treinamento; adicionalmente à seleção realizada no subconjunto de atributos, utilizados na divisão em cada nó das árvores (BREINMAN, 2001; BELGIU; DRĂGUŢ, 2016).

Ao construir um conjunto de árvores de decisão extremamente aleatório, o algoritmo Random Forest também evita a propagação de erros, produzindo modelos de dados mais estáveis (HUSSIEN et al., 2017).

Entre os algoritmos *white box* utilizados na pesquisa, o kNN demonstrou os melhores resultados em todos os tempos de predição, com diferenças não significantes em relação ao algoritmo com o maior coeficiente de correlação. Conquanto para a execução do algoritmo kNN seja necessário a definição da quantidade *k* de vizinhos mais próximos, o algoritmo não executa

processos intrínsecos de seleção de atributos e obteve resultados melhores sem associação com o algoritmo de seleção de atributos, em todos os intervalos de predição. Apesar de não existirem diferenças estatisticamente significantes, os resultados foram perceptivelmente melhores e, talvez, o algoritmo tenha se beneficiado da alta dimensionalidade dos conjuntos de dados.

Em relação às morbidades, existe uma grande diversidade dos aspectos envolvidos: tanto em relação aos processos de mineração de dados que foram executados, mas principalmente quanto às características clínicas das doenças. A utilização de uma abordagem de tabulação dos capítulos do CID-10 permitiu uma visão amplamente estruturada acerca da correlação entre os procedimentos ambulatoriais e hospitalares futuros, assim como a capacidade de predição dos modelos. Contudo, a utilização de dados da incidência de doenças do CID-10, ao invés do agrupamento em capítulos, pode conduzir a uma abordagem altamente detalhada, que também envolveria uma discussão sobre seus aspectos clínicos.

Não obstante, em razão das características de algumas das morbidades existentes, a predição de eventos de qualquer natureza – incluindo a incidência de internação hospitalar, que foi objeto dessa pesquisa – proporcionaria pouco relevância sob o aspecto da melhoria da capacidade de gerenciamento. A ocorrência de morbidades relacionadas a eventos como a gravidez, o parto e o puerpério é facilmente prevista em razão de seu aspecto temporal. Outrossim, morbidades relacionadas a causas externas não podem ser previstas através de modelagem de dados ambulatoriais, por serem eventos normalmente extemporâneos.

Apesar dos resultados apontarem correlação forte em todos os intervalos de tempo de predição, transtornos mentais e comportamentais é um grupo de morbidades que notadamente apresenta pouca relação com a ocorrência anterior de procedimentos ambulatoriais de diagnóstico em laboratório clínico. Outros aspectos podem estar relacionados com os resultados encontrados para esse capítulo do CID-10 e estudos mais detalhados, que considerem dados da ocorrência individual das doenças, podem permitir a evidenciação de informação importante.

Por outro lado, uma evidência de que os modelos de dados construídos nos processos de mineração de dados produziram bons resultados foi a correlação muito forte encontrada para o capítulo III do CID-10, que agrupa doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários. Grande parte dos procedimentos ambulatoriais de diagnóstico em laboratório clínico são exames laboratoriais sanguíneos, amplamente utilizados na investigação dessas doenças. Situação semelhante foi encontrada na correlação do capítulo que agrupa sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte (capítulo XVIII), que também contém códigos relacionados a sinais encontrados nos exames de diagnóstico em laboratório clínico executados em âmbito hospitalar.

A modelagem do conjunto de dados que continham as quantidades de internações hospitalares de doenças dos olhos e anexos (capítulo VII) e doenças da pele e do tecido subcutâneo (capítulo XII) também apresentaram resultados de correlação muito forte entre as predições realizadas e os valores reais. Assim como os outros capítulos do CID-10, diferentes doenças e condições de saúde estão agrupadas e, para permitir uma discussão mais aprofundada dos motivos que conduziram a esses resultados, seria necessário a execução de um projeto de mineração de dados específico. Desta forma, a condução desses experimentos pode produzir conhecimento importante a respeito da correlação obtida nesses grupos de doenças.

Outras morbidades de interesse são aquelas que, por estarem relacionadas com maiores taxas de mortalidade hospitalar, possuem importância para sua predição, não apenas relacionado ao gerenciamento de sistemas de saúde, como também sob o aspecto clínico. De acordo com os dados mais atualizados (BRASIL, 2019) – apresentados na Tabela 10 –, os cinco grupos de doenças que causaram as maiores taxas de mortalidade, respectivamente, são: doenças do aparelho circulatório, neoplasias, causas externas, doenças do aparelho respiratório e doenças endócrinas nutricionais e metabólicas. Dentre esses grupos de doenças, o capítulo Neoplasmas (tumores) possui correlação muito forte em ao menos um dos intervalos de tempo de predição (um mês). Desta forma, estudos futuros que considerem esse grupo e suas doenças individualmente, e que utilizem a mineração de dados para evidenciação de informação latente ou padrões desconhecidos, podem resultar na geração de conhecimento profícuo importante.

Tabela 10 - Mortalidade no Brasil em 2017, segundo capítulos do CID-10

Capítulo do CID-10	**Frequência**
Doenças do aparelho circulatório	358.882
Neoplasias (tumores)	221.821
Causas externas de morbidade e mortalidade	158.657
Doenças do aparelho respiratório	155.620
Doenças endócrinas nutricionais e metabólicas	79.662
Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	71.822
Doenças do aparelho digestivo	66.052
Algumas doenças infecciosas e parasitárias	54.874
Doenças do aparelho geniturinário	40.470
Doenças do sistema nervoso	38.786
Algumas afecções originadas no período perinatal	21.458
Transtornos mentais e comportamentais	12.858

Capítulo do CID-10	Frequência
Malformações congênitas, deformidades e anomalias cromossômicas	10.995
Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários	6.622
Doenças da pele e do tecido subcutâneo	6.100
Doenças do sistema osteomuscular e do tecido conjuntivo	5.912
Gravidez parto e puerpério	1.874
Doenças do ouvido e da apófise mastoide	179
Doenças do olho e anexos	19
Contatos com serviços de saúde	-
Códigos para propósitos especiais	-

Fonte: Sistema de Informações sobre Mortalidade - SIM (2017)

Os resultados obtidos demonstraram pouca diferença entres os coeficientes de correlação ρ de Pearson obtidos nos intervalos de um mês, três meses e seis meses. O resultado médio do coeficiente de correlação para o intervalo de tempo de três meses foi melhor entre os períodos analisados.

Efetivamente, diversos fatores podem afetar a definição de um adequado intervalo para predição, como o tempo de disponibilização dos dados, a capacidade e o tempo de resposta a um evento inesperado, bem como a relação custo e benefício para resposta a uma determinada circunstância.

Particularmente importante no centro cirúrgico de um hospital, o planejamento da capacidade de atendimento e a alocação de recursos utiliza as informações provenientes do *case-mix* nos diferentes níveis de decisão: estratégico, tático e operacional. O nível estratégico possui longo horizonte de planejamento, com objetivo de aprimorar a utilização dos recursos e a distribuição de orçamento, envolvendo problemas de médio e longo prazo e demandando informações e previsões. No entanto, os níveis tático e operacional envolvem decisões relacionadas a distribuição operacional de recursos humanos, estrutura física, equipamentos e insumos, em um horizonte de tempo de curto prazo, através do gerenciamento programático ou operacional do cronograma de disponibilidade dos recursos (KOPPKA et al., 2018; ZHU et al., 2018).

Desta forma, a aplicabilidade das previsões de morbidade hospitalar possui relevância nas atividades envolvidas no nível de decisão estratégico e, consequentemente, em um horizonte de médio a longo prazo. Ma e Demeulemeester (2013) declararam que as decisões

usualmente são realizadas a cada seis meses, intervalo de tempo que permitiria a adequada realocação de recursos quando necessário. Sob outra perspectiva, esse período também seria adequado para a eficiente alocação dos recursos necessários para atendimento às demandas, com o menor custo possível.

6.3 APLICAÇÕES

Os modelos preditivos produzidos através da mineração de dados podem ser aplicados em sistemas informatizados automatizados. Seus resultados podem ser exibidos quase em tempo real em um painel de gerenciamento clínico e auxiliar a tomada de decisões. As predições produzidas podem ser aplicadas como uma referência, em conjunto com outros indicadores comumente utilizados, para avaliar o desempenho de um sistema de saúde (GRAHAM et al., 2018).

E o monitoramento do desempenho é a chave para a sustentabilidade dos serviços de saúde. O aumento dos custos e da demanda por serviços de saúde estabeleceram uma nova perspectiva sobre a sustentabilidade, principalmente quando se trata do sistema de saúde financiado pelo governo. O monitoramento do desempenho está diretamente relacionado com a melhoria da eficiência na prestação de serviço, ainda muito prejudicada pela aleatoriedade da demanda de atenção à saúde. Kumar e Anjomshoa (2019) prospectam ser possível projetar um sistema, através da mineração de dados, que alcance uma utilização de mais de 90% (noventa por cento) dos recursos existentes; permitindo, assim, gerenciar sistema de saúde de forma mais eficiente, ao minimizar o efeito da aleatoriedade na demanda por serviços de saúde.

A confirmação da capacidade de geração de conhecimento através da mineração de dados, provenientes dos bancos de dados do sistema público de saúde brasileiro, e a proposição da utilização da informação para a melhoria da gestão de sistemas de saúde são as principais contribuições dessa pesquisa. Autores corroboram que a mineração de dados pode aperfeiçoar o gerenciamento de sistemas de saúde, tornando mais eficiente a alocação de recursos (AGRAVAL et al., 2016; GRAHAM et al., 2018; KAZEMI, MIRROSHANDEL, 2018; MEHTA; PANDIT, 2018; RICHTER; KHOSHGOFTAAR, 2018; SHIMODA et al., 2018; WIRATMADJA et al., 2018).

Em razão da natureza complexa e multidimensional das atividades hospitalares, seus processos de gestão necessitam de modelos de dados precisos, que permitam o planejamento adequado da alocação de recursos, principalmente em contextos de limitada disponibilidade orçamentária. Consequentemente, é importante compreender o perfil nosológico de um hospital

– também denominado de *case-mix* –, que consiste em grupos de pacientes que possuem um conjunto inter-relacionado de características, como por exemplo a gravidade da doença, o risco de morte, o prognóstico, a complexidade do tratamento, a urgência de intervenção e a quantidade de recursos necessários (HORNBROOK, 1982).

Os pacientes de um mesmo grupo demandam os mesmos recursos, como tempo igual na sala de cirurgia ou de internação. Determinar quantos pacientes de cada grupo de patologia podem ser tratados em um hospital possibilita a maximização da eficiência na alocação de recursos e o planejamento da capacidade do hospital. A aplicação equilibrada dos recursos também possui a propriedade de melhorar o nível de serviço de saúde ofertado ao paciente (MA; DEMEULEMEESTER, 2013; GRAHAM et al., 2018).

Prever a demanda de saúde em hospitais possibilita seu gerenciamento adequado, principalmente naquelas unidades em que esta pode ser maior que sua capacidade de atendimento. Desta forma, é possível ajustar a capacidade ou adotar ações corretivas alternativas. O gerenciamento adequado da demanda e oferta de serviços permite uma solução sistêmica para o gerenciamento hospitalar, com processos que alcancem a otimização do uso de recursos e a garantia do nível de serviço ofertado, garantindo o equilíbrio entre a qualidade dos cuidados de saúde e o custo da prestação dos serviços. Esse planejamento envolve a previsão de quantidade e de atributos específicos dos recursos necessários para atendimento adequado à demanda (SITEPU et al., 2018).

A capacidade de controlar e identificar fatores que afetam processo de tomada de decisão permite melhorar o gerenciamento clínico e de recursos, tornando-se excelente ferramenta de gestão hospitalar (WIRATMADJA et al., 2018).

Esses exemplos de aplicação podem ser prospectados restringindo-se apenas àqueles relacionados ao gerenciamento de sistemas de saúde. Quando esses limites são extrapolados, principalmente se consideradas aplicações clínicas, constrói-se a perspectiva de que a mineração de dados pode causar uma mudança comportamental em Saúde.

6.4 DESAFIOS PARA UTILIZAÇÃO DA MINERAÇÃO DE DADOS EM SAÚDE

Ravì et al. (2017) consideraram que a mineração de dados pode melhorar e desenvolver sistemas de apoio à tomada de decisão, permitindo o aprimoramento tanto na qualidade da atenção quanto na acessibilidade dos serviços de saúde. No entanto, os autores indicaram algumas razões que podem limitar essa capacidade, entre essas: a complexidade dos dados, em razão de estrutura variável; amostragem irregular e ausência de dados; quantidade de dados

muito grande, principalmente quando incluem imagens médicas, dados de sensores, resultados de laboratório e relatórios de texto não estruturados; dependência da longitudinalidade dos eventos clínicos e de diagnóstico e tratamento de doenças; incapacidade de abordagens tradicionais de aprendizado de máquina para modelar conjuntos de dados grandes e não estruturados; dificuldade de interpretação dos resultados, dificultando a adoção dos métodos no cenário clínico.

Nenhum modelo obtido por mineração de dados é um completo substituto para o julgamento humano, mas sim importante ferramenta que pode ser aplicada como suporte à decisão clínica. Profissionais de saúde devem compreender as limitações dos processos, pois a má utilização de algoritmos pode conduzir a efeitos adversos não esperados (WANG et al., 2018).

Consequentemente, infere-se duas vertentes para a consolidação da utilização da mineração de dados como ferramenta em gerenciamento de sistemas de saúde. O aprimoramento da mineração de dados, através da utilização de processos mais sofisticados nos procedimentos de mineração, bem como a pesquisa e desenvolvimento de algoritmos específicos para análise de dados de saúde podem melhorar os resultados dos modelos preditivos, contribuindo para fortalecer a utilização das aplicações. E, uma segunda vertente, está na disseminação da informação a profissionais de saúde sobre as possibilidades que a mineração de dados oferece para melhorar os processos em saúde, como também o aprimoramento e desenvolvimento profissional, incorporando a disciplina nos estudos acadêmicos nos ramos de Ciência da Saúde.

7 CONCLUSÃO

Através dos resultados obtidos na pesquisa foi possível concluir que a mineração de dados nos registros de procedimentos ambulatoriais e hospitalares permitiu a predição da morbidade hospitalar, através da aplicação de diferentes métodos de mineração de dados. Resultados diferentes foram alcançados dependendo do método aplicado, do grupo de doenças analisado e do intervalo de tempo de predição proposto. Consequentemente, projetos de mineração de dados que utilizem dados de um grupo específico ou das doenças individualmente devem ser objetos de estudos futuros.

O melhor intervalo de tempo para a predição da morbidade hospitalar foi de três meses, com o maior coeficiente de correlação ρ de Pearson de 0,5823. Esse resultado foi seguido pela obtenção de uma média de 0,5568, no intervalo de um mês, e 0,5414, no intervalo de seis meses.

No intervalo de tempo de predição de um mês, ao aplicar Random Forest, em conjunto com o algoritmo de seleção de atributos, foi obtido o melhor resultado médio entre os algoritmos utilizados (0,6493). No capítulo XVI do CID-10 (Algumas afecções originadas no período perinatal) o resultado médio alcançado foi o maior, com o coeficiente de correlação de 0,8725.

Considerando o intervalo de tempo de três meses, o método de mineração de dados que executou o algoritmo Random Forest associado a um algoritmo de seleção de atributos resultou na maior média do coeficiente de correlação (0,6806). O grupo de doenças com maior resultado médio, organizado por capítulos do CID-10, foi Algumas afecções originadas no período perinatal (capítulo XVI), que obteve o coeficiente de correlação 0,8801.

Nos testes realizados no intervalo de tempo de predição de seis meses, o algoritmo Random Forest obteve o maior resultado do coeficiente de correlação ρ de Pearson (0,6286). Mais uma vez, o capítulo XVI do CID-10 (Algumas afecções originadas no período perinatal) alcançou o melhor resultado, com o coeficiente de correlação de 0,8649.

Os resultados médios reportados sugeriram quais são os métodos de mineração de dados e grupos de doenças que permitem a obtenção de modelos de dados que alcançam bons resultados de predição. Contudo, os resultados individuais conduzem a uma avaliação melhor da capacidade de predição de morbidade hospitalar alcançado. Individualmente, o maior coeficiente de correlação ρ de Pearson foi obtido no intervalo de tempo de predição de três meses, através do método que aplicou o algoritmo Random Forest, associado com o algoritmo de seleção de atributos, no grupo de doenças do capítulo XVI do CID-10 (Algumas afecções originadas no período perinatal), com o resultado de 0,9682.

As predições da morbidade hospitalar obtidas podem minimizar o efeito indesejado da aleatoriedade da demanda por serviços de saúde no processo de tomada decisão. Predizer o fluxo de pacientes em unidades hospitalares proporciona uma eficiente alocação de recursos e um gerenciamento apropriado da capacidade instalada. O conhecimento gerado pela mineração de dados executada nessa pesquisa pode subsidiar o adequado planejamento na gestão hospitalar, conduzindo hospitais, públicos e privados, ao equilíbrio financeiro desejado e à melhora do nível de qualidade dos serviços prestados aos pacientes.

REFERÊNCIAS

DATASUS. Ministério da Saúde. **Disseminação de informações em saúde: Sistema de Informações Ambulatoriais do SUS (SIASUS)**. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=22&pad=31655. Acessado em: 27 out 2018.

DATASUS. Ministério da Saúde. **Tabulador de Dados para Windows**. Versão 4.1.3. Disponível em: http://www.datasus.gov.br/tabwin. Acessado em: 07 ago. 2018.

RAVÌ, D.; WONG, C.; DELIGIANNI, F.; BERTHELOT, M.; ANDREU-PEREZ, J.; LO, B.; YANG, G. Z. Deep learning for health informatics. IEEE journal of biomedical and health informatics, v. 21, n. 1, p. 4-21, 2016.

AFZAL, M.; HUSSAIN, M.; KHAN, W. A.; ALI, T.; LEE, S.; HUH, E.; AHMAD, H. F.; JAMSHED, A.; IQBAL, H.; IRFAN, M.; Hydari, M. A. Comprehensible knowledge model creation for cancer treatment decision making. **Computers in biology and medicine**, v. 82, p. 119-129, 2017.

AGRAWAL, A.; MATHIAS, J.; BAKER, D.; CHOUDHARY, A. Five year life expectancy calculator for older adults. In: **2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)**. IEEE, 2016. p. 1280-3.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine learning**, v. 6, n. 1, p. 37-66, 1991.

ALMADANI O.; ALSHAMMARI R. Prediction of stroke using data mining classification techniques. **International Journal of Advanced Computer Science and Applications**, v. 9, n. 1, p. 457-60, 2018.

ALSHAHRANI, M.; ZHU, F.; SAMEH, A.; ZHENG, L.; MUMTAZ, S. Evaluating the influence of Twitter on the Saudi Arabian stock market indicators. In: **5th International Symposium on Data Mining Applications**. Springer, 2018. p. 113-32.

ALTHUNAYAN, L.; ALSAHDI, N.; SYED, L. Comparative analysis of different classification algorithms for prediction of diabetes disease. In: **Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing**. ACM, 2017. p. 144.

ANGERMUELLER, C.; PÄRNAMAA, T.; PARTS, L.; STEGLE, O. Deep learning for computational biology. **Molecular systems biology**, v. 12, n. 7, 2016.

AWAD, A.; BADER-EL-DEN, M.; MCNICHOLAS, J.; BRIGGS, J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. **International journal of medical informatics**, v. 108, p. 185-95, 2017.

BARROS, E. F.; ROMÃO, W.; CONSTANTINO, A. A.; DE SOUZA, C. L. Pré-processamento para mineração de dados sobre beneficiários de planos de saúde suplementar. **Journal of Health Informatics**, v. 3, n. 1, 2011.

BAŞAR, M. D., AKAN, A. Chronic kidney disease prediction with reduced individual classifiers. **Journal of Electrical and Electronics Engineering**, v. 18, n. 2, p. 249-55, 2018.

BAURIN, N.; MORLEY, D.; OCHIAI, L.; GUERGOVA-KURAS, M.; AFSHAR, M.; COUDEVILLE, L. Management of dengue hospitalizations in Brazil during and outside epidemic periods: Insights from Data Mining. **American Journal of Tropical Medicine and Hygiene**, v. 95, n. 5, p. 67, nov. 2017.

BECKER, R. A.; CHAMBERS, J. M.; WILKS, A. R. **The S Language; A Programming Environment for Data Analysis and Graphics**. 1988.

BELGIU, M.; DRĂGUŢ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24-31, 2016.

BELLINGER, C.; MOHOMED JABBAR, M.S.; ZAÏANE, O.; OSORNIO-VARGAS, A. A systematic review of data mining and machine learning for air pollution epidemiology. **BMC Public Health**, v. 17, n. 1, 2017.

BIBAULT, J. E.; GIRAUD, P.; BURGUN, A. Big data and machine learning in radiation oncology: state of the art and future prospects. **Cancer letters**, v. 382, n. 1, p. 110-7, 2016.

BIRJALI, M.; BENI-HSSANE, A.; ERRITALI, M. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. **Procedia Computer Science**, v. 113, p. 65-72, 2017.

BORAH, M. S.; BHUYAN, B. P.; PATHAK, M. S.; BHATTACHARYA, P. K. Machine Learning in Predicting Hemoglobin Variants. **International Journal of Machine Learning and Computing**, v. 8, n. 2, 2018.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the fifth annual workshop on Computational learning theory**, ACM, 1992. p. 144-52.

BRASIL. Ministério da Saúde. Gabinete do Ministro. Portaria nº 2.848, de 6 de novembro de 2007. Publica a Tabela de Procedimentos, Medicamentos, Órteses, Próteses e Materiais Especiais – OPM do Sistema Único de Saúde. **Diário Oficial da União**, Brasília, DF, 07 nov. 2007.

BRASIL. Ministério da Saúde. Gabinete do Ministro. Portaria nº 3.462, de 11 de novembro de 2010. Estabelece critérios para alimentação dos Bancos de Dados Nacionais dos Sistemas de Informação da Atenção à Saúde. **Diário Oficial da União**, Brasília, DF, 12 nov. 2010.

BRASIL. Ministério da Saúde. **Tabulador de Dados para Windows**. Brasília: DATASUS, 2018. Versão 4.1.3. Disponível em: http://www.datasus.gov.br/tabwin. Acessado em: 07 ago. 2018.

BRASIL. Ministério da Saúde. **Disseminação de informações em saúde: Sistema de Informações Ambulatoriais do SUS (SIASUS)**. Brasília: DATASUS, 2018. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=22&pad=31655. Acessado em: 27 out 2018.

BRASIL. Ministério da Saúde. **SIM – Sistema de Informações de Mortalidade**. Brasília: DATASUS, 2019. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=060701. Acessado em: 02 a go. 2019.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CAMACHO, D. M.; COLLINS, K. M.; POWERS, R. K.; COSTELLO, J. C.; COLLINS, J. J. Next-generation machine learning for biological networks. **Cell**, v. 173, n. 7, p. 1581-92, 2018.

CASTALDO, R.; MELILLO, P.; IZZO, R.; DE LUCA, N.; PECCHIA, L. Fall prediction in hypertensive patients via short-term HRV Analysis. **IEEE journal of biomedical and health informatics**, v. 21, n. 2, p. 399-406, 2017.

CHAN, T. M.; LI, Y.; CHIAU, C. C.; ZHU, J.; JIANG, J.; HUO, Y. Imbalanced target prediction with pattern discovery on clinical data repositories. **BMC Medical Informatics and Decision Making**, v. 17, n. 1, p. 47, 2017.

CHICCO D. Ten quick tips for machine learning in computational biology. **BioData Mining**, v. 10, n. 1, dez. 2017.

CHING, T.; HIMMELSTEIN, D. S.; BEAULIEU-JONES, B. K.; KALININ, A. A.; DO, B. T.; WAY, G. P.; FERRERO, E.; AGAPOW, P. M.; ZIETZ, M.; HOFFMAN, M. M.; XIE, W.; ROSEN, G. L.; LENGERICH, B. J.; ISRAELI, J.; LANCHANTIN, J.; WOLOSZYNEK, S.; CARPENTER, A. E.; SHRIKUMAR, A.; XU, J.; COFER, E. M.; LAVENDER, C. A.; TURAGA, S. C.; ALEXANDARI, A. M.; LU, Z.; HARRIS, D. J.; DECAPRIO, D.; QI, Y.; KUNDAJE, A.; PENG, Y.; WILEY, L. K.; SEGLER, M. H. S.; BOCA, S. M.; SWAMIDASS, S. J.; HUANG, A.; GITTER, A.; GREENE, C. S. Opportunities and obstacles for deep learning in biology and medicine. **Journal of The Royal Society Interface**, v. 15, n. 141, p. 20170387, 2018.

CHOWDHURY, M. H.; ISLAM, M. K.; KHAN, S. I. Imputation of missing healthcare data. In: **2017 20th International Conference of Computer and Information Technology (ICCIT)**. IEEE, 2017. p. 1-6.

CIRKOVIC, B. A.; ISAILOVIC, V.; NIKOLIC, D.; SAVELJIC, I.; PARODI, O.; FILIPOVIC, N. Prediction of Coronary Plaque Progression Using Data Driven Approach. In: **International Conference on Future Access Enablers of Ubiquitous and Intelligent Infrastructures**. Springer, 2017. p. 227-33.

DAQQA, K. A. A.; MAGHARI, A. Y.; AL SARRAJ, W. F. Prediction and diagnosis of leukemia using classification algorithms. **Information Technology (ICIT), 2017 8th International Conference on**. IEEE, 2017. p. 638-43.

DRUCKER, H.; BURGES, C. J.; KAUFMAN, L.; SMOLA, A. J.; VAPNIK, V. Support vector regression machines. In: **Advances in neural information processing systems**. 1997. p. 155-61.

EL NAQA, I. Perspectives on making big data analytics work for oncology. **Methods**, v. 111, p. 32-44, 2016.

EVANS, J. D. **Straightforward statistics for the behavioral sciences**. Thomson Brooks/Cole Publishing Co, 1996.

FARZI, S.; KIANIAN, S.; RASTKHADIVE, I. Predicting serious diabetic complications using hidden pattern detection. In: **4th International Conference on Knowledge-Based Engineering and Innovation**. IEEE, 2018. p. 63-8.

FAWAZ, H. I.; FORESTIER, G.; WEBER, J.; IDOUMGHAR, L.; MULLER, P. A. Deep learning for time series classification: a review. **Data Mining and Knowledge Discovery**, v. 33, n. 4, p. 917-963, 2019.

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.

FERLAY, J.; COLOMBET, M.; SOERJOMATARAM, I.; MATHERS, C.; PARKIN, D. M.; PIÑEROS, M.; ZNAOR, A.; BRAY, F. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. **International journal of cancer**, v. 144, n. 8, p. 1941-53, 2018.

FREEMAN, N.; ZHAO, M.; MELOUK, S. An iterative approach for case mix planning under uncertainty. **Omega**, v. 76, p. 160-73, 2018.

FREIRE, S. M.; SOUZA, R. C.; ALMEIDA, R. T. Integrating Brazilian health information systems in order to support the building of data warehouses. **Biomededical Engineering**, v. 31, n. 3, p. 196-207, 2015.

GIACALONE, M.; CUSATELLI, C.; SANTARCANGELO, V. Big Data Compliance for Innovative Clinical Models. **Big Data Research**, v. 12, p. 35-40, 2018.

GILLIES, R. J.; KINAHAN, P. E.; HRICAK, H. Radiomics: images are more than pictures, they are data. **Radiology**, v. 278, n. 2, p. 563-77, 2015.

GONZÁLEZ, J.; FERRER, J. C.; CATALDO, A.; ROJAS, L. A proactive transfer policy for critical patient flow management. **Health care management science**, v. 22, n. 2, p. 287-303, 2019.

GRAHAM, B.; BOND, R.; QUINN, M.; MULVENNA, M. Using Data Mining to Predict Hospital Admissions From the Emergency Department. **IEEE Access**, v. 6, p. 10458-69, 2018.

GUPTA, N.; AHUJA, N.; MALHOTRA, S.; BALA, A.; KAUR, G. Intelligent heart disease prediction in cloud environment through ensembling. **Expert Systems**, v. 34, n. 3, 2017.

HORNBROOK, M. C. Hospital case mix: its definition, measurement and use: Part I. The conceptual framework. **Medical care review**, v. 39, n. 1, p. 1-43, 1982.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International Journal of Data Mining & Knowledge Management Process**, v. 5, n. 2, p. 1, 2015.

HUNTA, S.; YOOYATIVONG, T.; AUNSRI, N. A novel integrated action crossing method for drug-drug interaction prediction in non-communicable diseases. **Computer methods and programs in biomedicine**, v. 163, p. 183-93, 2018.

HUSSIEN, S. O.; ELKHATEM, S. S.; OSMAN, N.; IBRAHIM, A. O. A review of data mining techniques for diagnosing hepatitis. In: **2017 Sudan Conference on Computer Science and Information Technology (SCCSIT)**. IEEE, 2017. p. 1-6.

JHA, S. K.; PAN, Z.; ELAHI, E.; PATEL, N. A comprehensive search for expert classification methods in disease diagnosis and prediction. **Expert Systems**, v. 36, n. 1, p. e12343, 2019.

KALAISELVI, K.; SUJARANI, P. Correlation Feature Selection (CFS) and Probabilistic Neural Network (PNN) for diabetes disease prediction International. **Journal of Engineering and Technology**, 2018.

KAUR, P.; KUMAR, R.; KUMAR, M. A healthcare monitoring system using random forest and internet of things (IoT). **Multimedia Tools and Applications**, p. 1-12, 2019.

KAZEMI, Y.; MIRROSHANDEL, S. A. A novel method for predicting kidney stone type using ensemble learning. **Artificial intelligence in medicine**, v. 84, p. 117-126, 2018.

KOPPKA, L.; WIESCHE, L.; SCHACHT, M.; WERNERS, B. Optimal distribution of operating hours over operating rooms using probabilities. **European Journal of Operational Research**, v. 267, n. 3, p. 1156-71, 2018.

KUMAR, A.; ANJOMSHOA, H. A Two-Stage Model to Predict Surgical Patients' Lengths of Stay From an Electronic Patient Database. **IEEE journal of biomedical and health informatics**, v. 23, n. 2, p. 848-56, 2019.

KUMAR, A.; ANJOMSHOA, H. A Two-Stage Model to Predict Surgical Patients' Lengths of Stay From an Electronic Patient Database. **IEEE journal of biomedical and health informatics**, v. 23, n. 2, p. 848-56, 2018.

KUMAR, N.; KHATRI, S. Implementing WEKA for medical data classification and early disease prediction. In: **2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)**. IEEE, 2017. p. 1-6.

KUNJIR, A.; SAWANT, H.; SHAIKH, N. F. Data mining and visualization for prediction of multiple diseases in healthcare. In: **2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)**. IEEE, 2017. p. 329-34.

LEAVELL, H. R.; CLARK, E. G. **Medicina preventiva**. 1976.

LI, G.; ZHOU, X.; LIU, J.; CHEN, Y.; ZHANG, H.; CHEN, Y.; NIE, S. Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province. **PLoS neglected tropical diseases**, v. 12, n. 2, 2018.

LINDSAY, W. D.; AHERN, C. A.; TOBIAS, J. S.; BERLIND, C. G.; CHINNIAH, C.; GABRIEL, P. E.; GEE, J. C.; SIMONE 2ND, C. B. Automated data extraction and ensemble methods for predictive modeling of breast cancer outcomes after radiation therapy. **Medical physics**, v. 46, n. 2, p. 1054-1063, 2019.

LIU, C. L.; HSAIO, W. H.; TU, Y. C. Time series classification with multivariate convolutional neural network. **IEEE Transactions on Industrial Electronics**, v. 66, n. 6, p. 4788-4797, 2019.

LUO, L.; LIU, C.; FENG, L.; ZHAO, S.; GONG, R. A random forest and simulation approach for scheduling operation rooms: Elective surgery cancelation in a Chinese hospital urology department. **The International journal of health planning and management**, v. 33, n. 4, p. 941-66, 2018.

MA, G.; DEMEULEMEESTER, E. A multilevel integrative approach to hospital case mix and capacity planning. **Computers & Operations Research**, v. 40, n. 9, p. 2198-207, 2013.

MARTÍNEZ-GARCÍA, M.; SALINAS-ORTEGA, M.; ESTRADA-ARRIAGA, I.; HERNÁNDEZ-LEMUS, E.; GARCÍA-HERRERA, R.; VALLEJO, M. A systematic approach to analyze the social determinants of cardiovascular disease. **PloS one**, v. 13, n. 1, 2018.

MEHTA, N.; PANDIT, A. Concurrence of big data analytics and healthcare: A systematic review. **International journal of medical informatics**, v. 114, p. 57-65, 2018.

MIOTTO, R.; LI, L.; KIDD, B. A.; DUDLEY, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. **Scientific reports**, v. 6, p. 26094, 2016.

MLAKAR M.; PUDDU P. E.; SOMRAK M.; BONFIGLIO S.; LUŠTREK M. Mining telemonitored physiological data and patient-reported outcomes of congestive heart failure patients. **PLoS ONE**, v. 13, n. 3, mar. 2018.

MOSES, H.; MATHESON, D. H.; DORSEY, E. R.; GEORGE, B. P.; SADOFF, D.; YOSHIMURA, S. The anatomy of health care in the United States. **Jama**, v. 310, n. 18, p. 1947-1964, 2013.

NAVAZ A. N.; MOHAMMED E.; SERHANI M. A.; ZAKI N. The use of data mining techniques to predict mortality and length of stay in an ICU. In: **2016 12th International Conference on Innovations in Information Technology (IIT)**. IEEE, 2016. p. 1-5.

PELÁNEK, R. Metrics for evaluation of student models. **Journal of Educational Data Mining**, v. 7, n. 2, p. 1-19, 2015.

PIRES, F. A. **Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde**. 2011. Tese (Doutorado em Ciências) – Universidade de São Paulo. São Paulo.

QUDSI, D. H.; KARTIWI, M.; SALEH, N. B. Predictive Data Mining of Chronic Diseases Using Decision Tree: A Case Study of Health Insurance Company in Indonesia. **International Journal of Applied Engineering Research**, v. 12, n. 7, p. 1334-9, 2017.

QUINLAN, J. R. Learning with continuous classes. **5th Australian joint conference on artificial intelligence**, v. 92, p. 343-8, 1992.

RAU, H. H.; HSU, C. Y.; LIN, Y. A.; ATIQUE, S.; FUAD, A.; WEI, L. M.; HSU, M. H. Development of a web-based liver cancer prediction model for type II diabetes patients by

using an artificial neural network. **Computer methods and programs in biomedicine**, v. 125, p. 58-65, 2016.

RAVÌ, D.; WONG, C.; DELIGIANNI, F.; BERTHELOT, M.; ANDREU-PEREZ, J.; LO, B.; YANG, G. Z. Deep learning for health informatics. **IEEE journal of biomedical and health informatics**, v. 21, n. 1, p. 4-21, 2016.

RICHTER A. N.; KHOSHGOFTAAR T. M. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. **Artificial Intelligence in Medicine**, v. 90, p. 1-14, ago. 2018.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, v. 65, n. 6, p. 386, 1958.

ROTH, A.; SUBRAMANIAN, S.; GANAPATHIRAJU, M. K. Towards extracting supporting information about predicted protein-protein interactions. **IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)**, v. 15, n. 4, p. 1239-46, 2018.

SACRAMENTO, I. L. S. **Redução dimensional dos dados de entrada em previsões de consumo industrial de energia no longo prazo**. 2014. Dissertação (Mestrado em Ciências da Computação) – Universidade Federal de Santa Catarina. Florianópolis.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, v. 23, n. 19, p. 2507-2517, 2007.

SAHNI, N.; SIMON, G.; ARORA, R. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. **Journal of general internal medicine**, v. 33, n. 6, p. 921-8, 2018.

SANEJA, B.; RANI, R. A scalable correlation-based approach for outlier detection in wireless body sensor networks. **International Journal of Communication Systems**, v. 32, n. 7, p. e3918, 2019.

SANTOS, A. B. V.; CARVALHO, D. R. Predictive models for infant mortality in the state of Paraná. **Iberoamerican Journal of Applied Computing**, v. 7, n. 2, 2018.

SANTOS, R. S.; GUTIERREZ, M. A. Minersus - Ambiente computacional para extração de informações para a gestão da saúde pública por meio da mineração dos dados do SUS. **Research on Biomedical Engineering**, v. 24, n. 2, p. 77-90, 2011.

SHEARER, C. The CRISP-DM model: the new blueprint for data mining. **Journal of data warehousing**, v. 5, n. 4, p. 13-22, 2000.

SHI, Y.; LI, P.; YU, X.; WANG, H.; NIU, L. Evaluating Doctor Performance: Ordinal Regression-Based Approach. **Journal of medical Internet research**, v. 20, n. 7, 2018.

SHICKEL, B.; TIGHE, P. J.; BIHORAC, A.; RASHIDI, P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. **IEEE journal of biomedical and health informatics**, v. 22, n. 5, p. 1589-604, 2018.

SHIMODA, A.; ICHIKAWA, D.; OYAMA, H. Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. **Computer methods and programs in biomedicine, v.** 163, p. 39-46, 2018.

SITEPU, S.; MAWENGKANG, H.; HUSEIN, I. Optimization Model for Capacity Management and Bed Scheduling for Hospital. In: **IOP Conference Series: Materials Science and Engineering**. IOP Publishing, 2018. p. 012016.

SRIVASTAVA, A.; MAHMOOD, A.; SRIVASTAVA, R. A Comparative Analysis of SVM Random Forest Methods for Protein Function Prediction. In: **2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)**. IEEE, 2017. p. 1008-10.

TAZIN, N.; SABAB, S. A.; CHOWDHURY, M. T. Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. In: **2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)**. IEEE, 2016. p. 1-6.

TESFAYE, B.; ATIQUE, S.; ELIAS, N.; DIBABA, L.; SHABBIR, S. A.; KEBEDE, M. Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach. **Computer methods and programs in biomedicine**, v. 140, p. 45-51, 2017.

VAN DER PLOEG, T.; AUSTIN, P. C.; STEYERBERG, E. W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. **BMC medical research methodology**, v. 14, n. 1, p. 137, 2014.

VAPNIK, V. N. **The nature of Statistical learning theory**. Springer science & business mediaNew York: Wiley-Interscience Publication, 2013.

VAUGHN, D. A.; VAN DEEN, W. K.; KERR, W.T.; MEYER, T. R.; BERTOZZI, A. L.; HOMMES, D. W.; COHEN, M. S. Using insurance claims to predict and improve hospitalizations and biologics use in members with inflammatory bowel diseases. **Journal of Biomedical Informatics**, v. 81, p. 93-101, 2018.

VENABLES, W. N.; SMITH, D. M. **An introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics**. 2009. Disponível em: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf. Acessado em: 27 out 2018.

VIANNA, R. C. X. F.; MORO, C. M. C. D. B.; MOYSÉS, S. J.; CARVALHO, D.; NIEVOLA, J. C. Mineração de dados e características da mortalidade infantil. **Cadernos de Saúde Pública**, n. 26, p. 535-542, 2010.

WAGSTAFF, K. Machine learning that matters. **arXiv preprint arXiv:1206.4656**, 2012.

WANG, Y.; KUNG, L.; WANG, W. Y. C.; CEGIELSKI, C. G. An integrated big data analytics-enabled transformation model: Application to health care. **Information and Management**, v. 55, n. 1, p. 64-79, 2018.

WERBOS, P. **Beyond Regression:" New Tools for Prediction and Analysis in the Behavioral Sciences**. Dissertação (Ph. D. in Statistics) – Harvard University. Cambridge. 1974.

WIRATMADJA, I. I.; SALAMAH, S. Y.; GOVINDARAJU, R. Healthcare Data Mining: Predicting Hospital Length of Stay of Dengue Patients. **Journal of Engineering and Technological Sciences**, v. 50, n. 1, p. 110-126, 2018.

WITTEN, I. H.; FRANK, E.; HALL, M. A. The WEKA workbench. **Data mining: Practical machine learning tools and techniques**. 4. ed. Morgan Kaufmann, 2016.

WORLD HEALTH ORGANIZATION. **Classifications**. Suíça, 2019. Disponível em: https://www.who.int/classifications/icd/en/. Acessado em: 12 jul. 2019.

ZHAO, Y. **R and data mining: Examples and case studies**. Academic Press, 2012.

ZHU, S.; FAN, W.; YANG, S.; PEI, J.; PARDALOS, P. M. Operating room planning and surgical case scheduling: a review of literature. **Journal of Combinatorial Optimization**, v. 37, n. 3, p. 757-805, 2019.

APÊNDICE A – Testes estatísticos realizados em linguagem R, nos resultados dos processos de mineração dos dados para predição com intervalo de um mês.

Quadro 1 - Testes de normalidade Shapiro-Wilk no grupo de resultados médios por algoritmo (intervalo de um mês).

	ibk	ibk_atributos
statistic	0.8638184	0.8619913
p.value	0.006018898	0.005573467
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	m5	m5_atributos
statistic	0.8883008	0.8730896
p.value	0.01746436	0.00894068
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	rf	rf_atributos
statistic	0.8821368	0.8554647
p.value	0.01327389	0.004247298
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	rna	rna_atributos
statistic	0.9344291	0.8371866
p.value	0.1517224	0.002031605
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	svm	svm_atributos
statistic	0.8916231	0.8697953
p.value	0.02028202	0.007759674
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"

Fonte: Elaborado pelo autor (2019).

Quadro 2 - Testes de normalidade Shapiro-Wilk no grupo de resultados médios por capítulos do CID-10 (intervalo de um mês).

	capitulo_01	capitulo_02
statistic	0.5831056	0.4977785
p.value	3.480649e-05	3.425116e-06
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_03	capitulo_04
statistic	0.5196971	0.4933232
p.value	6.194485e-06	3.037221e-06
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_05	capitulo_06
statistic	0.485437	0.9504521
p.value	2.455669e-06	0.6738517
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_07	capitulo_08
statistic	0.5036976	0.9725711
p.value	4.018638e-06	0.9136064
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_09	capitulo_10
statistic	0.9618277	0.6749033
p.value	0.8064966	0.0004379217
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"

```
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_11                   capitulo_12
statistic 0.6452052                     0.6024841
p.value   0.0001921952                  5.920778e-05
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_13                   capitulo_14
statistic 0.9159064                     0.8275539
p.value   0.3240727                     0.03126499
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_15                   capitulo_16
statistic 0.8702374                     0.5143893
p.value   0.1006046                     5.365468e-06
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_17                   capitulo_18
statistic 0.9573123                     0.5439261
p.value   0.754846                      1.19542e-05
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_19                   capitulo_20
statistic 0.5369652                     0.6631272
p.value   9.894124e-06                  0.0003157695
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_21                   capitulo_22
statistic 0.7060639                     0.8839624
p.value   0.001043367                   0.1448552
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
```

Fonte: Elaborado pelo autor (2019).

Quadro 3 - Testes de Kruskal-Wallis (intervalo de um mês).

```
        Kruskal-Wallis rank sum test

data:  dados by algoritmo
Kruskal-Wallis chi-squared = 38.368, df = 9, p-value = 1.497e-05

        Kruskal-Wallis rank sum test

data:  dados by capitulo
Kruskal-Wallis chi-squared = 165.39, df = 21, p-value < 2.2e-16
```

Fonte: Elaborado pelo autor (2019).

Quadro 4 - Teste de Wilcoxon de comparações múltiplas dos resultados médios por algoritmos (intervalo de um mês).

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  dados and algoritmo

              ibk     ibk_atributos m5      m5_atributos rf      rf_atributos
ibk_atributos 1.00000 -             -       -            -       -
m5            1.00000 1.00000       -       -            -       -
m5_atributos  1.00000 1.00000       1.00000 -            -       -
rf            1.00000 1.00000       1.00000 1.00000      -       -
rf_atributos  1.00000 1.00000       1.00000 1.00000      1.00000 -
rna           0.00036 0.00160       0.00024 0.00036      1.5e-05 3.5e-05
```

```
rna_atributos 1.00000 1.00000       1.00000 1.00000      1.00000 1.00000
svm           1.00000 1.00000       1.00000 1.00000      1.00000 1.00000
svm_atributos 1.00000 1.00000       1.00000 1.00000      1.00000 1.00000
              rna     rna_atributos svm
ibk_atributos -       -             -
m5            -       -             -
m5_atributos  -       -             -
rf            -       -             -
rf_atributos  -       -             -
rna           -       -             -
rna_atributos 0.00039 -             -
svm           0.00014 1.00000       -
svm_atributos 5.6e-05 1.00000       1.00000

P value adjustment method: holm
```

Fonte: Elaborado pelo autor (2019).

Quadro 5 - Teste de Wilcoxon de comparações múltiplas dos resultados médios por capítulo do CID-10 (intervalo de um mês).

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  dados and capitulo

            capitulo_01 capitulo_02 capitulo_03 capitulo_04 capitulo_05
capitulo_02 0.1491      -           -           -           -
capitulo_03 0.1630      1.0000      -           -           -
capitulo_04 1.0000      0.1630      0.2102      -           -
capitulo_05 0.7492      0.9757      1.0000      1.0000      -
capitulo_06 0.0025      0.0025      0.0025      0.0044      0.0044
capitulo_07 0.1491      1.0000      1.0000      0.1491      0.6553
capitulo_08 0.0562      0.0044      0.0044      0.0799      0.1131
capitulo_09 0.0562      0.0044      0.0044      0.0799      0.0799
capitulo_10 1.0000      0.1491      0.1491      0.6553      0.1491
capitulo_11 0.1491      0.1491      0.1491      0.1491      0.1491
capitulo_12 0.2604      1.0000      1.0000      0.6553      1.0000
capitulo_13 0.0142      0.0025      0.0025      0.0799      0.0799
capitulo_14 0.1491      0.1131      0.1131      0.1491      0.1491
capitulo_15 0.1491      0.0368      0.0368      0.1491      0.1491
capitulo_16 0.1491      0.6553      0.5358      0.1491      0.2604
capitulo_17 0.0562      0.0025      0.0025      0.1131      0.1131
capitulo_18 0.1491      1.0000      1.0000      0.1491      0.5358
capitulo_19 0.2102      0.7492      0.7492      0.2102      0.6553
capitulo_20 1.0000      1.0000      1.0000      1.0000      1.0000
capitulo_21 1.0000      0.1491      0.1491      1.0000      0.3274
capitulo_22 0.1491      0.1491      0.1131      0.1491      0.1491
            capitulo_06 capitulo_07 capitulo_08 capitulo_09 capitulo_10
capitulo_02 -           -           -           -           -
capitulo_03 -           -           -           -           -
capitulo_04 -           -           -           -           -
capitulo_05 -           -           -           -           -
capitulo_06 -           -           -           -           -
capitulo_07 0.0025      -           -           -           -
capitulo_08 0.0240      0.0025      -           -           -
capitulo_09 0.0240      0.0025      1.0000      -           -
capitulo_10 0.0044      0.1491      0.1491      0.1131      -
capitulo_11 0.0044      0.1491      0.1491      0.1131      0.7492
capitulo_12 0.0025      1.0000      0.0044      0.0044      0.1491
capitulo_13 0.0799      0.0025      1.0000      1.0000      0.1131
capitulo_14 0.0044      0.0562      0.1491      0.1630      0.2604
capitulo_15 0.0240      0.0240      1.0000      1.0000      0.1491
```

capitulo_16	0.0025	0.6553	0.0025	0.0025	0.1491
capitulo_17	0.0142	0.0025	1.0000	1.0000	0.1491
capitulo_18	0.0025	1.0000	0.0025	0.0025	0.1491
capitulo_19	0.0025	0.8345	0.0044	0.0044	0.1491
capitulo_20	0.1131	1.0000	0.1491	0.1491	1.0000
capitulo_21	0.0025	0.1491	0.0368	0.0562	1.0000
capitulo_22	0.0044	0.0799	0.1491	0.1491	0.2604
	capitulo_11	capitulo_12	capitulo_13	capitulo_14	capitulo_15
capitulo_02	-	-	-	-	-
capitulo_03	-	-	-	-	-
capitulo_04	-	-	-	-	-
capitulo_05	-	-	-	-	-
capitulo_06	-	-	-	-	-
capitulo_07	-	-	-	-	-
capitulo_08	-	-	-	-	-
capitulo_09	-	-	-	-	-
capitulo_10	-	-	-	-	-
capitulo_11	-	-	-	-	-
capitulo_12	0.1491	-	-	-	-
capitulo_13	0.1491	0.0025	-	-	-
capitulo_14	0.1491	0.1131	0.1491	-	-
capitulo_15	0.1491	0.0368	0.2102	1.0000	-
capitulo_16	0.1491	0.7492	0.0025	0.0562	0.0240
capitulo_17	0.1491	0.0025	1.0000	0.1491	0.8345
capitulo_18	0.1491	1.0000	0.0025	0.0562	0.0240
capitulo_19	0.1491	0.8345	0.0025	0.1131	0.0368
capitulo_20	0.3274	1.0000	0.1491	0.1491	0.1491
capitulo_21	0.7492	0.2102	0.0081	0.1630	0.1491
capitulo_22	0.1491	0.1491	0.0799	1.0000	1.0000
	capitulo_16	capitulo_17	capitulo_18	capitulo_19	capitulo_20
capitulo_02	-	-	-	-	-
capitulo_03	-	-	-	-	-
capitulo_04	-	-	-	-	-
capitulo_05	-	-	-	-	-
capitulo_06	-	-	-	-	-
capitulo_07	-	-	-	-	-
capitulo_08	-	-	-	-	-
capitulo_09	-	-	-	-	-
capitulo_10	-	-	-	-	-
capitulo_11	-	-	-	-	-
capitulo_12	-	-	-	-	-
capitulo_13	-	-	-	-	-
capitulo_14	-	-	-	-	-
capitulo_15	-	-	-	-	-
capitulo_16	-	-	-	-	-
capitulo_17	0.0025	-	-	-	-
capitulo_18	1.0000	0.0025	-	-	-
capitulo_19	1.0000	0.0025	1.0000	-	-
capitulo_20	0.2102	0.1491	0.5358	0.4173	-
capitulo_21	0.1491	0.0368	0.1491	0.1491	1.0000
capitulo_22	0.0562	0.1131	0.0562	0.1131	0.1491
	capitulo_21				
capitulo_02	-				
capitulo_03	-				
capitulo_04	-				
capitulo_05	-				
capitulo_06	-				
capitulo_07	-				
capitulo_08	-				
capitulo_09	-				
capitulo_10	-				
capitulo_11	-				
capitulo_12	-				

```
capitulo_13 -
capitulo_14 -
capitulo_15 -
capitulo_16 -
capitulo_17 -
capitulo_18 -
capitulo_19 -
capitulo_20 -
capitulo_21 -
capitulo_22 0.1630

P value adjustment method: holm
```

Fonte: Elaborado pelo autor (2019).

APÊNDICE B – Testes estatísticos realizados em linguagem R, nos resultados dos processos de mineração dos dados para predição com intervalo de três meses.

Quadro 6 - Testes de normalidade Shapiro-Wilk no grupo de resultados médios por algoritmo (intervalo de três meses).

	ibk	ibk_atributos
statistic	0.8843887	0.9078529
p.value	0.01466636	0.04280657
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	m5	m5_atributos
statistic	0.9200697	0.9073472
p.value	0.07627179	0.04180582
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	rf	rf_atributos
statistic	0.8814499	0.8741262
p.value	0.01287753	0.009350604
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	rna	rna_atributos
statistic	0.9333717	0.9306676
p.value	0.1442287	0.1266894
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	svm	svm_atributos
statistic	0.8942325	0.9145358
p.value	0.022829	0.05863403
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"

Fonte: Elaborado pelo autor (2019).

Quadro 7 - Testes de normalidade Shapiro-Wilk no grupo de resultados médios por capítulos do CID-10 (intervalo de três meses)

	capitulo_01	capitulo_02
statistic	0.5879936	0.4854226
p.value	3.979098e-05	2.454717e-06
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_03	capitulo_04
statistic	0.4515716	0.5097701
p.value	9.886109e-07	4.735312e-06
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_05	capitulo_06
statistic	0.558049	0.9297834
p.value	1.755795e-05	0.4457655
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_07	capitulo_08
statistic	0.5927759	0.8063918
p.value	4.536267e-05	0.01733629
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_09	capitulo_10
statistic	0.8415229	0.7026072
p.value	0.04601675	0.0009473916
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"

```
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_11                   capitulo_12
statistic 0.6501661                     0.4728564
p.value   0.000220478                   1.750435e-06
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_13                   capitulo_14
statistic 0.8852543                     0.7653936
p.value   0.1498466                     0.005495705
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_15                   capitulo_16
statistic 0.743668                      0.5091911
p.value   0.002988214                   4.661765e-06
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_17                   capitulo_18
statistic 0.8347285                     0.5597505
p.value   0.03814378                    1.83914e-05
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_19                   capitulo_20
statistic 0.4190042                     0.6025686
p.value   4.138443e-07                  5.934534e-05
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_21                   capitulo_22
statistic 0.7004606                     0.7503307
p.value   0.000892307                   0.003601935
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
```

Fonte: Elaborado pelo autor (2019).

Quadro 8 - Testes de Kruskal-Wallis (intervalo de três meses).

```
        Kruskal-Wallis rank sum test

data:  dados by algoritmo
Kruskal-Wallis chi-squared = 47.099, df = 9, p-value = 3.761e-07

        Kruskal-Wallis rank sum test

data:  dados by capitulo
Kruskal-Wallis chi-squared = 154.88, df = 21, p-value < 2.2e-16
```

Fonte: Elaborado pelo autor (2019).

Quadro 9 - Teste de Wilcoxon de comparações múltiplas dos resultados médios por algoritmos (intervalo de três meses).

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  dados and algoritmo

              ibk     ibk_atributos m5      m5_atributos rf      rf_atributos
ibk_atributos 1       -             -       -            -       -
m5            1       1             -       -            -       -
m5_atributos  1       1             1       -            -       -
rf            1       1             1       1            -       -
rf_atributos  1       1             1       1            1       -
rna           2.5e-06 7.4e-06       3.1e-07 3.0e-06      6.3e-06 4.9e-07
```

```
rna_atributos 1       1             1       1            1       1
svm           1       1             1       1            1       1
svm_atributos 1       1             1       1            1       1
              rna     rna_atributos svm
ibk_atributos -       -             -
m5            -       -             -
m5_atributos  -       -             -
rf            -       -             -
rf_atributos  -       -             -
rna           -       -             -
rna_atributos 1.4e-06 -             -
svm           6.6e-05 1             -
svm_atributos 9.3e-07 1             1

P value adjustment method: holm
```

Fonte: Elaborado pelo autor (2019).

Quadro 10 - Teste de Wilcoxon de comparações múltiplas dos resultados médios por capítulo do CID-10 (intervalo de três meses).

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  dados and capitulo

            capitulo_01 capitulo_02 capitulo_03 capitulo_04 capitulo_05
capitulo_02 0.1764      -           -           -           -
capitulo_03 0.1764      1.0000      -           -           -
capitulo_04 1.0000      0.3222      0.1764      -           -
capitulo_05 0.9086      1.0000      0.1764      1.0000      -
capitulo_06 0.0093      0.0025      0.0025      0.0157      0.0093
capitulo_07 0.1987      1.0000      1.0000      1.0000      1.0000
capitulo_08 0.1764      0.1349      0.0930      0.1764      0.1764
capitulo_09 0.1764      0.0930      0.0643      0.1764      0.1349
capitulo_10 1.0000      0.1987      0.1764      1.0000      1.0000
capitulo_11 0.1764      0.1764      0.1764      0.1764      0.1764
capitulo_12 0.1764      1.0000      1.0000      0.1764      0.1764
capitulo_13 0.1764      0.1349      0.0643      0.1764      0.1349
capitulo_14 0.1764      0.1764      0.0643      0.1764      0.1764
capitulo_15 1.0000      0.1764      0.1764      0.1987      0.1987
capitulo_16 0.1764      0.1764      0.1764      0.1764      0.1764
capitulo_17 0.1764      0.1349      0.1349      0.1764      0.1764
capitulo_18 0.1764      1.0000      1.0000      0.1764      0.1764
capitulo_19 0.1764      0.1764      0.1764      0.1764      0.1764
capitulo_20 1.0000      1.0000      1.0000      1.0000      1.0000
capitulo_21 1.0000      0.1764      0.1764      1.0000      1.0000
capitulo_22 0.1764      0.1764      0.1764      0.1764      0.1764
            capitulo_06 capitulo_07 capitulo_08 capitulo_09 capitulo_10
capitulo_02 -           -           -           -           -
capitulo_03 -           -           -           -           -
capitulo_04 -           -           -           -           -
capitulo_05 -           -           -           -           -
capitulo_06 -           -           -           -           -
capitulo_07 0.0025      -           -           -           -
capitulo_08 0.0093      0.0930      -           -           -
capitulo_09 0.0413      0.0643      1.0000      -           -
capitulo_10 0.0025      0.3222      0.1764      0.1349      -
capitulo_11 0.0093      0.1764      0.1764      0.1349      0.1987
capitulo_12 0.0025      1.0000      0.1349      0.0930      0.1764
capitulo_13 0.0413      0.0643      1.0000      1.0000      0.1349
capitulo_14 0.0093      0.0930      1.0000      1.0000      0.1764
capitulo_15 0.0093      0.1764      0.1764      0.1349      1.0000
```

capitulo_16	0.0025	0.1764	0.0047	0.0047	0.1764
capitulo_17	0.0157	0.1349	1.0000	1.0000	0.1764
capitulo_18	0.0025	0.6208	0.0025	0.0025	0.1764
capitulo_19	0.0025	0.1764	0.0047	0.0047	0.1764
capitulo_20	0.1764	1.0000	0.1764	0.1764	1.0000
capitulo_21	0.0025	0.3222	0.1764	0.1349	1.0000
capitulo_22	0.0025	0.1764	0.1764	0.1349	0.1764
	capitulo_11	capitulo_12	capitulo_13	capitulo_14	capitulo_15
capitulo_02	-	-	-	-	-
capitulo_03	-	-	-	-	-
capitulo_04	-	-	-	-	-
capitulo_05	-	-	-	-	-
capitulo_06	-	-	-	-	-
capitulo_07	-	-	-	-	-
capitulo_08	-	-	-	-	-
capitulo_09	-	-	-	-	-
capitulo_10	-	-	-	-	-
capitulo_11	-	-	-	-	-
capitulo_12	0.1764	-	-	-	-
capitulo_13	0.1349	0.1349	-	-	-
capitulo_14	0.1764	0.1764	1.0000	-	-
capitulo_15	1.0000	0.1764	0.1764	0.1764	-
capitulo_16	0.1764	0.6208	0.0157	0.0093	0.1764
capitulo_17	0.6208	0.1349	1.0000	1.0000	0.3222
capitulo_18	0.1764	1.0000	0.0157	0.0047	0.1764
capitulo_19	0.1764	0.1764	0.0157	0.0157	0.1764
capitulo_20	0.1764	0.3222	0.1764	0.1764	0.4912
capitulo_21	0.6208	0.1764	0.1349	0.1764	1.0000
capitulo_22	1.0000	0.1764	0.1987	0.1764	0.7345
	capitulo_16	capitulo_17	capitulo_18	capitulo_19	capitulo_20
capitulo_02	-	-	-	-	-
capitulo_03	-	-	-	-	-
capitulo_04	-	-	-	-	-
capitulo_05	-	-	-	-	-
capitulo_06	-	-	-	-	-
capitulo_07	-	-	-	-	-
capitulo_08	-	-	-	-	-
capitulo_09	-	-	-	-	-
capitulo_10	-	-	-	-	-
capitulo_11	-	-	-	-	-
capitulo_12	-	-	-	-	-
capitulo_13	-	-	-	-	-
capitulo_14	-	-	-	-	-
capitulo_15	-	-	-	-	-
capitulo_16	-	-	-	-	-
capitulo_17	0.0643	-	-	-	-
capitulo_18	1.0000	0.0413	-	-	-
capitulo_19	1.0000	0.0643	1.0000	-	-
capitulo_20	0.1764	0.1764	0.3222	0.1764	-
capitulo_21	0.1764	0.2448	0.1764	0.1764	1.0000
capitulo_22	0.1764	1.0000	0.1764	0.1764	0.1764
	capitulo_21				
capitulo_02	-				
capitulo_03	-				
capitulo_04	-				
capitulo_05	-				
capitulo_06	-				
capitulo_07	-				
capitulo_08	-				
capitulo_09	-				
capitulo_10	-				
capitulo_11	-				
capitulo_12	-				

```
capitulo_13 -
capitulo_14 -
capitulo_15 -
capitulo_16 -
capitulo_17 -
capitulo_18 -
capitulo_19 -
capitulo_20 -
capitulo_21 -
capitulo_22 0.3831

P value adjustment method: holm
```

Fonte: Elaborado pelo autor (2019).

APÊNDICE C – Testes estatísticos realizados em linguagem R, nos resultados dos processos de mineração dos dados para predição com intervalo de seis meses.

Quadro 11 - Testes de normalidade Shapiro-Wilk no grupo de resultados médios por algoritmo (intervalo de seis meses).

	ibk	ibk_atributos
statistic	0.8958533	0.8879211
p.value	0.0245785	0.01716963
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	m5	m5_atributos
statistic	0.9115454	0.874696
p.value	0.05091137	0.009584322
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	rf	rf_atributos
statistic	0.8978856	0.864462
p.value	0.02697307	0.006184675
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	rna	rna_atributos
statistic	0.935469	0.863672
p.value	0.1594648	0.00598186
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"
	svm	svm_atributos
statistic	0.8872748	0.8716884
p.value	0.01667996	0.008416629
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[algoritmo == x]"	"dados[algoritmo == x]"

Fonte: Elaborado pelo autor (2019).

Quadro 12 - Testes de normalidade Shapiro-Wilk no grupo de resultados médios por capítulos do CID-10 (intervalo de seis meses).

	capitulo_01	capitulo_02
statistic	0.5784342	0.5230141
p.value	3.063085e-05	6.776828e-06
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_03	capitulo_04
statistic	0.4616614	0.4872144
p.value	1.295827e-06	2.576113e-06
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_05	capitulo_06
statistic	0.5356347	0.9392817
p.value	9.543041e-06	0.5450492
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_07	capitulo_08
statistic	0.5835286	0.9187604
p.value	3.521181e-05	0.3467319
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"
data.name	"dados[capitulo == x]"	"dados[capitulo == x]"
	capitulo_09	capitulo_10
statistic	0.9014676	0.6687636
p.value	0.22736	0.000369245
method	"Shapiro-Wilk normality test"	"Shapiro-Wilk normality test"

```
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_11                   capitulo_12
statistic 0.6336869                     0.5043651
p.value   0.0001397983                  4.09175e-06
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_13                   capitulo_14
statistic 0.8392089                     0.776791
p.value   0.04317151                    0.007566135
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_15                   capitulo_16
statistic 0.7454967                     0.4723321
p.value   0.003145384                   1.725937e-06
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_17                   capitulo_18
statistic 0.8757252                     0.5095128
p.value   0.1165001                     4.702489e-06
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_19                   capitulo_20
statistic 0.469227                      0.6105125
p.value   1.587751e-06                  7.382292e-05
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
          capitulo_21                   capitulo_22
statistic 0.8188695                     0.6751883
p.value   0.0245566                     0.0004414049
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "dados[capitulo == x]"        "dados[capitulo == x]"
```

Fonte: Elaborado pelo autor (2019).

Quadro 13 - Testes de Kruskal-Wallis (intervalo de seis meses).

```
        Kruskal-Wallis rank sum test

data:  dados by algoritmo
Kruskal-Wallis chi-squared = 42.437, df = 9, p-value = 2.734e-06

        Kruskal-Wallis rank sum test

data:  dados by capitulo
Kruskal-Wallis chi-squared = 163.31, df = 21, p-value < 2.2e-16
```

Fonte: Elaborado pelo autor (2019).

Quadro 14 - Teste de Wilcoxon de comparações múltiplas dos resultados médios por algoritmos (intervalo de seis meses).

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  dados and algoritmo

              ibk     ibk_atributos m5      m5_atributos rf      rf_atributos
ibk_atributos 1.00000 -             -       -            -       -
m5            1.00000 1.00000       -       -            -       -
m5_atributos  1.00000 1.00000       1.00000 -            -       -
rf            1.00000 1.00000       1.00000 1.00000      -       -
rf_atributos  1.00000 1.00000       1.00000 1.00000      1.00000 -
rna           2.3e-06 0.00044       4.7e-06 9.6e-06      2.3e-06 1.3e-05
```

```
rna_atributos 1.00000 1.00000       1.00000 1.00000      1.00000 1.00000
svm           1.00000 1.00000       1.00000 1.00000      1.00000 1.00000
svm_atributos 1.00000 1.00000       1.00000 1.00000      1.00000 1.00000
              rna     rna_atributos svm
ibk_atributos -       -             -
m5            -       -             -
m5_atributos  -       -             -
rf            -       -             -
rf_atributos  -       -             -
rna           -       -             -
rna_atributos 6.8e-05 -             -
svm           5.9e-05 1.00000       -
svm_atributos 3.2e-06 1.00000       1.00000

P value adjustment method: holm
```

Fonte: Elaborado pelo autor (2019).

Quadro 15 - Teste de Wilcoxon de comparações múltiplas dos resultados médios por capítulo do CID-10 (intervalo de seis meses).

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  dados and capitulo

            capitulo_01 capitulo_02 capitulo_03 capitulo_04 capitulo_05
capitulo_02 0.1728      -           -           -           -
capitulo_03 0.1691      1.0000      -           -           -
capitulo_04 1.0000      0.2060      0.1691      -           -
capitulo_05 1.0000      0.5729      0.1691      1.0000      -
capitulo_06 0.0045      0.0025      0.0025      0.0896      0.0045
capitulo_07 1.0000      1.0000      0.2060      1.0000      1.0000
capitulo_08 0.0896      0.0025      0.0087      0.1276      0.0087
capitulo_09 0.1691      0.0025      0.0251      0.1691      0.0896
capitulo_10 0.4108      0.1691      0.1691      0.1691      0.1691
capitulo_11 0.1691      0.1691      0.1691      0.1691      0.1691
capitulo_12 0.1691      1.0000      1.0000      0.1691      0.3215
capitulo_13 0.0387      0.0025      0.0025      0.1691      0.0025
capitulo_14 0.1691      0.0149      0.1276      0.1691      0.1691
capitulo_15 0.1691      0.1276      0.1691      0.1691      0.1691
capitulo_16 0.1691      0.1691      0.1728      0.1691      0.1691
capitulo_17 0.1691      0.0025      0.0896      0.1691      0.1276
capitulo_18 0.1691      0.2060      1.0000      0.1691      0.1691
capitulo_19 0.1691      0.1728      0.4713      0.1691      0.1691
capitulo_20 1.0000      1.0000      0.6490      1.0000      1.0000
capitulo_21 0.5729      0.1691      0.1691      0.1691      0.1691
capitulo_22 0.1691      0.1691      0.1691      0.1691      0.1691
            capitulo_06 capitulo_07 capitulo_08 capitulo_09 capitulo_10
capitulo_02 -           -           -           -           -
capitulo_03 -           -           -           -           -
capitulo_04 -           -           -           -           -
capitulo_05 -           -           -           -           -
capitulo_06 -           -           -           -           -
capitulo_07 0.0045      -           -           -           -
capitulo_08 1.0000      0.0087      -           -           -
capitulo_09 0.1728      0.0601      0.7434      -           -
capitulo_10 0.0896      0.1691      0.1691      0.1691      -
capitulo_11 0.0251      0.1691      0.0896      0.1691      0.5729
capitulo_12 0.0045      0.7434      0.0087      0.0896      0.1691
capitulo_13 1.0000      0.0025      1.0000      0.1728      0.1691
capitulo_14 0.1691      0.1276      0.1728      1.0000      0.1691
capitulo_15 0.1276      0.1691      0.1691      0.2060      0.1691
```

capitulo_16	0.0025	0.1691	0.0025	0.0087	0.1691
capitulo_17	0.1276	0.1276	0.6490	1.0000	0.1691
capitulo_18	0.0025	0.1691	0.0025	0.0025	0.1691
capitulo_19	0.0025	0.1691	0.0025	0.0149	0.1691
capitulo_20	0.1691	1.0000	0.1691	0.1691	0.4108
capitulo_21	0.0025	0.1691	0.0025	0.0251	1.0000
capitulo_22	0.0251	0.1691	0.1276	0.1691	0.1691
	capitulo_11	capitulo_12	capitulo_13	capitulo_14	capitulo_15
capitulo_02	-	-	-	-	-
capitulo_03	-	-	-	-	-
capitulo_04	-	-	-	-	-
capitulo_05	-	-	-	-	-
capitulo_06	-	-	-	-	-
capitulo_07	-	-	-	-	-
capitulo_08	-	-	-	-	-
capitulo_09	-	-	-	-	-
capitulo_10	-	-	-	-	-
capitulo_11	-	-	-	-	-
capitulo_12	0.1691	-	-	-	-
capitulo_13	0.0387	0.0025	-	-	-
capitulo_14	0.1691	0.1691	0.1691	-	-
capitulo_15	0.5729	0.1691	0.1691	0.4713	-
capitulo_16	0.1691	0.2060	0.0025	0.0896	0.1276
capitulo_17	0.1691	0.1276	0.1691	1.0000	0.4108
capitulo_18	0.1691	1.0000	0.0025	0.0045	0.1276
capitulo_19	0.1691	0.2494	0.0025	0.1276	0.1276
capitulo_20	0.1691	1.0000	0.1691	0.1691	0.1691
capitulo_21	0.6490	0.1691	0.0025	0.1276	0.1691
capitulo_22	1.0000	0.1691	0.0387	0.1691	1.0000
	capitulo_16	capitulo_17	capitulo_18	capitulo_19	capitulo_20
capitulo_02	-	-	-	-	-
capitulo_03	-	-	-	-	-
capitulo_04	-	-	-	-	-
capitulo_05	-	-	-	-	-
capitulo_06	-	-	-	-	-
capitulo_07	-	-	-	-	-
capitulo_08	-	-	-	-	-
capitulo_09	-	-	-	-	-
capitulo_10	-	-	-	-	-
capitulo_11	-	-	-	-	-
capitulo_12	-	-	-	-	-
capitulo_13	-	-	-	-	-
capitulo_14	-	-	-	-	-
capitulo_15	-	-	-	-	-
capitulo_16	-	-	-	-	-
capitulo_17	0.0149	-	-	-	-
capitulo_18	1.0000	0.0025	-	-	-
capitulo_19	1.0000	0.0149	1.0000	-	-
capitulo_20	0.1691	0.1691	0.1728	0.1691	-
capitulo_21	0.1691	0.0251	0.1691	0.1691	0.4108
capitulo_22	0.1691	0.1691	0.1691	0.1691	0.1691
	capitulo_21				
capitulo_02	-				
capitulo_03	-				
capitulo_04	-				
capitulo_05	-				
capitulo_06	-				
capitulo_07	-				
capitulo_08	-				
capitulo_09	-				
capitulo_10	-				
capitulo_11	-				
capitulo_12	-				

```
capitulo_13 -
capitulo_14 -
capitulo_15 -
capitulo_16 -
capitulo_17 -
capitulo_18 -
capitulo_19 -
capitulo_20 -
capitulo_21 -
capitulo_22 0.4108

P value adjustment method: holm
```

Fonte: Elaborado pelo autor (2019).