

UNIVERSIDADE DE SÃO PAULO
HOSPITAL DE REABILITAÇÃO DE ANOMALIAS CRANIOFACIAS

PATRICK PEDREIRA SILVA

**Application of data mining to support knowledge discovery
in the context of the Randomized Clinical Trial - Florida
Project**

**Aplicação da mineração de dados para apoio à descoberta
de conhecimento no contexto do Estudo Clínico
Randomizado - Projeto Florida**

BAURU

2021

PATRICK PEDREIRA SILVA

**Application of data mining to support knowledge discovery
in the context of the Randomized Clinical Trial - Florida
Project**

**Aplicação da mineração de dados para apoio à descoberta
de conhecimento no contexto do Estudo Clínico
Randomizado - Projeto Florida**

Tese constituída por artigo apresentada ao Programa de Pós-graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais da Universidade de São Paulo para obtenção do título de Doutor em Ciências, na área de concentração Fissuras Orofaciais e Anomalias Relacionadas.

Orientadora: Profa. Dra. Jeniffer de Cássia Rillo Dutka

Co-orientador: Prof. Dr. Elvio Gilberto da Silva

Versão corrigida

BAURU

2021

UNIVERSIDADE DE SÃO PAULO
HOSPITAL DE REABILITAÇÃO DE ANOMALIAS CRANIOFACIAS

R. Silvio Marchione, 3-20

Caixa Postal: 1501

17012-900 - Bauru - SP – Brasil

Prof. Dr. Vahan Agopyan - Reitor da USP

Prf. Dr. Carlos Ferreira dos Santos - Superintendente do HRAC/USP

Autorizo, exclusivamente para fins acadêmicos e científicos, a reprodução total ou parcial desta dissertação/tese, por processos fotocopiadores e outros meios eletrônicos.

Assinatura:



Pedreira Silva, Patrick

Aplicação da mineração de dados para apoio à descoberta de conhecimento no contexto do estudo clínico randomizado-Projeto Florida / Patrick Pedreira Silva. -- Bauru, 2021.

000 p.108 : il. ; 31 cm.

Tese (doutorado) -- Hospital de Reabilitação de Anomalias Craniofaciais, Universidade de São Paulo, 2021.

Comitê de Ética do HRAC-USP

Protocolo nº: 1.753.467

ERRATA

FOLHA DE APROVAÇÃO

Patrick Pedreira Silva

Tese apresentada ao Hospital de Reabilitação de Anomalias Craniofaciais da Universidade de São Paulo para obtenção do título de Doutor.

Área de concentração: Fissuras Orofaciais e Anomalias Relacionadas.

Aprovado em:

Banca Examinadora

Prof.(a) Dr.(a) _____

Instituição: _____

Prof.(a) Dr.(a) _____

Instituição: _____

Prof.(a) Dr.(a)

Instituição (Orientador(a))

Prof.(a) Dr.(a)

Presidente da Comissão de Pós-Graduação do HRAC-USP

Data de depósito da tese junto à SPG: ____/____/____

DEDICATÓRIA

Dedico este trabalho à minha família pelo apoio incondicional durante toda a minha vida, permitindo que eu realizasse sonhos como este do doutorado. Nominalmente, à minha mãe (Vera) pela sabedoria, apoio emocional e incentivo aos estudos, ao meu pai (Gilberto - *in memoriam*) que foi um pai maravilhoso e um grande apoiador, dando suporte para que seus filhos tivessem sempre o melhor possível, à minha irmã (Pollyanna) por ser minha primeira grande inspiração referente aos estudos e por ser sempre muito presente e, por fim, ao meu sobrinho (Antônio Carlos) que entrou na minha vida em 2010 e me deu a oportunidade de ser um humano melhor, exercendo o papel de tio.

À minha família bauruense (por consideração), nominalmente, a Flavia, a Elaine, o Flavio e o Lucas, que desde 2002 me acolheram como um membro da família, me ajudaram e deram suporte para que eu pudesse escrever minha história em Bauru e no estado de São Paulo.

Aos meus amigos (em especial, Elvio e Vinicius) que proporcionaram momentos de descontração e descanso mental frente aos estresses inerentes ao mundo acadêmico.

Ao meu pequeno pet Luck, cuja companhia e brincadeiras “fora de hora” foram fundamentais em muitos momentos.

A Deus pela força, pelo dom da vida e por colocar pessoas maravilhosas no meu caminho.

AGRADECIMENTOS

Em especial, à professora doutora Jeniffer Dutka, minha orientadora, pela oportunidade ao abrir as portas do doutorado e por todo acolhimento, competência, parceria e apoio durante todo o processo de elaboração deste trabalho.

À demais pesquisadores do HRAC pela parceria e ajuda na análise dos dados, escrita dos artigos e submissões às revistas e eventos.

Às meninas do Programa de Pós-Graduação pela ajuda nos assuntos burocráticos.

“Há um tempo em que é preciso abandonar as roupas usadas, que já têm a forma do nosso corpo e esquecer os nossos caminhos, que nos levam sempre aos mesmos lugares. É o tempo da travessia: e, se não ousarmos fazê-la, teremos ficado, para sempre, à margem de nós mesmos”.

Fernando Teixeira de Andrade

RESUMO

A quantidade de usuários da internet em todo o mundo vem aumentando de forma exponencial, alcançando cerca de 5,2 bilhões em 2021. Do mesmo modo, há um aumento significativo da velocidade de processamento e capacidade de memória dos computadores. Maior capacidade de processamento associada ao crescente aumento do número de usuários destas tecnologias têm produzido um problema de superabundância de dados, pois a capacidade de coletar e armazenar dados tem superado a habilidade de analisar e extrair conhecimento destes. A capacidade de extrair conhecimento útil e oculto nessa grande quantidade de dados e de agir com base nesse conhecimento está se tornando cada vez mais importante. Hospitais, clínicas e instituições ligados à área de saúde, de um modo geral, também ampliaram de forma significativa suas bases de dados, gerando impactos diretos na Saúde Pública uma vez que essa massa de dados fornece meios de subsidiar mecanismos de controle, procedimentos e, sobretudo, estudos clínicos que se transformem em benefícios para a sociedade. A Ciência da Computação apresenta um conjunto de técnicas e ferramentas destinadas à produção de informação relevante e à descoberta de conhecimentos em bases de dados de maneira inteligente e automática, por meio da chamada mineração de dados. Diante deste cenário, o objetivo geral deste trabalho foi aplicar técnicas de computação na área de ciências da saúde, sobretudo no escopo das fissuras labiopalatinas, buscando otimizar processos relacionados às práticas dessas ciências, por meio da aplicação de conceitos relacionados à inteligência artificial, ao aprendizado de máquina e à mineração de dados no contexto do HRAC/USP (Hospital de Reabilitação de Anomalias Craniofaciais – Universidade de São Paulo). Mais especificamente, a hipótese desta investigação é de que é possível identificar correlações e padrões de dados que permitam fornecer insights ao tratamento das fissuras labiopalatinas. Deste modo, foram propostas duas tarefas de pesquisa: a) comparação de algoritmos de aprendizado de máquina para a predição de ocorrência de fístulas após a palatoplastia primária em pacientes com fissura transforame unilateral (FTU) e b) utilização de técnicas de mineração de dados para a descoberta de conhecimento sobre fatores associados à ocorrência de fístulas após a palatoplastia primária. A análise dos dados revelou que a ausência de alguns sintomas (febre,

tosse, infecção) bem como características associadas à cirurgia em si (cirurgião, técnica, retalho de vômer) e ao paciente (hipernasalidade e sinais sugestivos de disfunção velofaríngea), podem ajudar a prever o sucesso ou insucesso da palatoplastia. Também revelou que fatores associados às complicações pós-operatórias (infecção, vômito, tosse e febre) bem como às características associadas à cirurgia em si (duração, técnicas cirúrgicas, retalho de Vômer, incisão relaxante) e ao paciente (idade na época da palatoplastia), podem ajudar a prever o sucesso ou insucesso da palatoplastia com relação à ocorrência de fístulas consideradas complicações. Quando se considera a capacidade de predição (correta e do maior número de casos) o algoritmo de melhor desempenho foi o de Máquina de Vetores-Suporte (SVM), cuja métrica de f-measure dentro da classe de insucessos foi a mais alta. Além das contribuições pontuais desta pesquisa salientadas nos artigos que compõem esta tese, podem ser evidenciadas outras contribuições potenciais deste trabalho, principalmente, a ampliação das parcerias entre as áreas de ciência da computação e ciências da saúde, sobretudo no escopo das fissuras labiopalatinas, o que permite otimizar processos relacionados às práticas dessas ciências, por meio das tecnologias da informação e comunicação, fomentando projetos multidisciplinares.

Palavras-chave: Fissuras. Fístulas. Inteligência Artificial. Mineração de Dados.

ABSTRACT

The number of internet users around the world has been increasing exponentially, reaching about 5.2 billion in 2021. In the same way, there is a significant increase in the processing speed and memory capacity of computers. Greater processing power associated with the increasing number of users of these technologies have produced a problem of data overabundance, as the ability to collect and store data has surpassed the ability to analyze and extract knowledge from it. The ability to extract useful and hidden knowledge from this vast amount of data and to act on that knowledge is becoming increasingly important. Hospitals, clinics and institutions linked to the healthcare area, in general, also have been expanding the way to form their databases, generating direct impacts on Public Health, since this mass of data offers means to subsidize mechanisms of control, procedures and above all, foster clinical studies that turn into benefits for society. Computer Science presents a set of techniques and tools aimed at production of relevant information and discovery of knowledge in databases in an intelligent and automatic way, through the so-called data mining. Given this scenario, the general objective of this work was to apply computational techniques in health sciences area, especially in the scope of cleft lip and palate, seeking to optimize processes related to the practices of these sciences, through the application of concepts related to artificial intelligence, to machine learning and data mining in the context of the HRAC/USP (Hospital for the Rehabilitation of Craniofacial Anomalies – University of São Paulo). More specifically, the hypothesis of this investigation is that it is possible to identify correlations and patterns in the data that provide insights into the treatment of cleft lip and palate. Thus, two research tasks were proposed in this project: a) comparison of machine learning algorithms for predicting the occurrence of fistulas after primary palatoplasty in patients with unilateral transforame cleft (UTC) and b) the use of data mining techniques for discovery of knowledge about factors associated with the occurrence of fistulas after primary palatoplasty. Data analysis revealed that the absence of some symptoms (fever, cough, infection) as well as characteristics associated with the surgery itself (surgeon, technique, vomer flap) and the patient (hypernasality and signs suggestive of velopharyngeal dysfunction), can help predict the success or failure of palatoplasty. It also revealed factors associated with postoperative

complications (infection, vomiting, cough and fever) as well as characteristics associated with the surgery itself (duration, surgical techniques, vomer flap, relaxing incision) and with the patient (age at the time of palatoplasty), can help predict the success or failure of palatoplasty in relation to the occurrence of fistulas considered complications. When considering the prediction capacity (correct and the largest number of cases), the best performing algorithm was the Support Vector Machine (SVM), whose f-measure metric within the failure class was the highest. In addition to the specific contributions of this research highlighted in the articles that make up this thesis, other potential contributions of this work can be evidenced, especially, by the expansion of partnerships between the areas of computer science and health sciences, particularly in the scope of cleft lip and palate, which allows optimizing processes related to the practices of these sciences, contributing with information and communication technologies, that foster future multidisciplinary projects.

Keywords: Cleft. Fistula. Artificial Intelligence. Data Mining.

LISTA DE ABREVIATURAS

AD	Árvore de Decisão
CBIS	Computer Based Information System
CSV	Comma-Separated Value
ECR-PF	Estudo Clínico Randomizado-Projeto Florida
e-health	eletronic health
FLP	Fissura Labiopalatina
FTU	Fissura Transforame Unilateral
HRAC/USP	Hospital de Reabilitação de Anomalias Craniofaciais – Universidade de São Paulo
IA	Inteligência Artificial
KDD	Knowledge Discovery in Databases (Descoberta de Conhecimento em bases de dados)
K-NN	K-Nearest Neighbor (K vizinhos mais próximos)
LDA	Linear Discriminant Analysis
MLP	Multilayer Perceptron
MTS	Multivariate Time Series
NBG	Naive Bayes Gaussiano
OMS	Organização Mundial da Saúde
RF	Random Forest (Floresta Aleatória)
RL	Regressão Logística
SVM	Support Vector Machine (Máquina de Vetores Suporte)
UTI	Unidade de Terapia Intensiva

SUMÁRIO

1	INTRODUÇÃO.....	16
2	REFERENCIAL TEÓRICO	25
2.1	INTELIGÊNCIA ARTIFICIAL, APRENDIZAGEM AUTOMÁTICA E MINERAÇÃO DE DADOS.....	25
2.2	PARADIGMAS DE APRENDIZAGEM E TAREFAS DA MINERAÇÃO.....	33
2.3	A MINERAÇÃO DE DADOS NA ÁREA DE SAÚDE.....	37
3	OBJETIVOS.....	46
4	RELAÇÃO ENTRE OS ARTIGOS.....	49
5	ARTIGOS.....	54
5.1	ARTIGO 1.....	54
	PREDIÇÃO DA OCORRÊNCIA DE FÍSTULAS NAS FISSURAS LABIOPALATINAS UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA.....	54
5.2	ARTIGO 2.....	70
	APLICAÇÃO DE MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO SOBRE FATORES ASSOCIADOS À OCORRÊNCIA DE FÍSTULAS APÓS PALATOPLASTIA	70
6	CONCLUSÃO GERAL.....	89
	REFERÊNCIAS.....	93
	ANEXO A – PARECER DO COMITÊ DE ÉTICA.....	102
	APÊNDICE A - DECLARAÇÃO DE USO EXCLUSIVO DE ARTIGOS EM DISSERTAÇÃO/TESE.....	107

1

Introdução

1 INTRODUÇÃO

A quantidade de usuários da internet em todo o mundo vem aumentando de forma exponencial, passando de 16 milhões em 1995 para cerca de 5,2 bilhões em 2021 (WORLD, 2021). Do mesmo modo, há um aumento significativo do número de componentes dos circuitos integrados que está diretamente relacionado à velocidade de processamento e capacidade de memória dos computadores. Maior capacidade de processamento associada ao crescente aumento do número de usuários destas tecnologias,

“[...] têm produzido um problema de superabundância de dados, pois a capacidade de coletar e armazenar dados tem superado a habilidade de analisar e extrair conhecimento destes”. (FERRARI e CASTRO, 2016, pag. 3).

A necessidade de entender grandes e complexos conjuntos de dados (o chamado, *big data*) aumentou em variados campos de tecnologia, negócios e ciências. *Big data* se refere a este grande volume de dados que são gerados universalmente a cada instante e que está expondo uma nova onda de tecnologia e arquitetura destinada a extrair valor desses dados, de modo a transformá-los em informações importantes e valiosas. Com esse grande aumento, a capacidade de extrair conhecimento útil e oculto nessa grande quantidade de dados e de agir com base nesse conhecimento está se tornando cada vez mais importante no mundo competitivo de hoje. Esse fenômeno pode ser observado em todos os campos e não apenas na internet já que os dados vêm sendo coletados e acumulados em um ritmo cada vez maior (KORTH, SILBERSCHATZ, SUDARSHAN, 2006; HEUSER, 2008; DATE, 2004). Segundo Minor (2017) 150 exabytes ou 10^{18} bytes de novos dados de saúde são gerados anualmente somente nos Estados Unidos (crescimento anual de 48%). Hospitais, clínicas e instituições ligadas à área de saúde, de um modo geral, também têm ampliado de forma significativa suas bases de dados, gerando impactos diretos na Saúde Pública.

A Saúde Pública pode ser definida como

“a arte e a ciência de prevenir a doença, prolongar a vida, promover a saúde e a eficiência física e mental mediante o esforço organizado da comunidade, abrangendo [...] o desenvolvimento de uma estrutura social que assegure a cada indivíduo na sociedade um padrão de vida adequado à manutenção da saúde”. (WINSLOW, 1920 apud VISELTEAR, 1982, p. 146).

Neste sentido, a Saúde Pública pauta-se no controle, na redução e na prevenção de doenças, assim como na promoção e manutenção da saúde de toda a população. Como forma de viabilizar seu funcionamento e fornecer meios de subsidiar mecanismos de controle, procedimentos e, sobretudo, estudos clínicos que se transformem em benefícios para a sociedade, torna-se imprescindível a geração de um grande volume de informações, que podem ser vistas como matérias-primas para a realização destas ações. Entretanto, esta explosão no volume de dados digitais, que ultrapassa a velocidade humana de interpretar e assimilar a informação, gerou a necessidade do desenvolvimento de novas técnicas e ferramentas que pudessem, além de processar os dados, permitir sua análise, de maneira inteligente e automática, para a descoberta de informações úteis e com aplicabilidade em diversos segmentos da sociedade (FAYYAD, 1998). Nesse sentido, a Ciência da Computação apresenta um conjunto de técnicas e ferramentas destinadas à produção de informação relevante e à descoberta de conhecimentos em bases de dados.

Um dos principais aspectos do processo de produção de informação envolve a aplicação de métodos específicos da chamada “mineração de dados” para descoberta e extração de padrões (AGRAWAL e PSAILA, 1995; KANTARDZIC, 2019). Muitos setores usam com êxito a mineração de dados (do inglês *data mining*): ela pode ser usada para ajudar o setor de varejo a modelar o comportamento do cliente, permite aos bancos prever a lucratividade dos usuários do sistema e apresenta usos semelhantes em telecomunicações, manufatura, indústria, ensino superior dentre outras.

A definição mais básica de mineração de dados é a análise de grandes conjuntos de dados para descobrir padrões e usá-los para prever a probabilidade de eventos futuros (FERRARI e CASTRO, 2016). O processo de aplicação de sistemas de informação baseados em computador (CBIS - *Computer Based Information System*) para descobrir o conhecimento dos dados já vem sendo usado como base da mineração de dados há alguns anos (VLAHOS; FERRATT; KNOEPFLE, 2004).

Particularmente a área do conhecimento da informática em saúde ou saúde eletrônica (*e-health*) tem apresentado um crescimento notável nas últimas décadas (SABBATINI, 2004). Segundo Marin e colaboradores (1990), a informática

contribui para a área de saúde por meio da utilização de recursos tecnológicos para o armazenamento e a manipulação dos dados, facilitando o acesso à informação, auxiliando, assim, a prática nos serviços de saúde. Os sistemas de computação permitem armazenar uma quantidade considerável de dados em bases eletrônicas (bancos de dados) e produzir uma grande quantidade de informação, porém, há a necessidade e o desafio de capacitar profissionais aptos a identificar e implementar ferramentas adequadas para manipular a informação disponível, agilizando a busca do conhecimento necessário aos objetivos da Saúde Pública.

Observa-se que nas últimas décadas a maioria das operações e atividades das instituições de saúde, sejam elas privadas ou públicas, tem sido registrada computacionalmente, acarretando grandes bases de dados eletrônicas. Ao mesmo tempo que estas instituições despendem tempo e esforço para construir e manter as suas bases de dados, na maioria das vezes, o conhecimento potencial contido nestas bases é subvalorizado ou subutilizado (FERRARI e CASTRO, 2016). O aprimoramento constante dos métodos de investigação e análise de dados dessas instituições, bem como da formação de profissionais com visão interdisciplinar e multidisciplinar favorece os estudos científicos. Neste cenário, a mineração de dados apresenta-se como uma alternativa eficaz para extração não trivial, a partir de grandes bases de dados, de conhecimento implícito, previamente desconhecido e potencialmente útil (padrões, relações, regras e correlação entre dados). Este processo apresenta a grande vantagem de ocorrer de modo automático ou semiautomático (FERRARI e CASTRO, 2016).

Um proeminente campo de pesquisa para a extração de informações em bases de dados é denominado KDD (*Knowledge Discovery in Databases*) ou, em português, Descoberta do Conhecimento em Bases de Dados. Este processo visa identificar padrões ou modelos que representem informação válida, inédita, potencialmente útil e essencialmente compreensível em uma coleção de dados (FAYYAD, 1998). Em geral, técnicas de mineração de dados possuem um papel preponderante nesse processo. A mineração de dados por meio de um conjunto de técnicas automáticas para a exploração em grandes massas de dados, objetiva encontrar novos padrões, tendências e relações. Através dessas técnicas, são possíveis a extração de informações úteis, a correlação de padrões existentes nos

dados e a descoberta de novos conhecimentos (os quais são dificilmente percebidos pelo ser humano).

A junção das abordagens de Estatística e Inteligência Artificial (IA) (áreas bastante consolidadas), fazem com que o *data mining* seja cada vez mais aceito e aplicado no meio científico (DHILLON et al., 2021). Os principais objetivos do *data mining* são a previsão e a descrição. Na previsão, variáveis conhecidas na base de dados são utilizadas para prever valores desconhecidos. Já a descrição é voltada para a busca de padrões que descrevam os dados de forma compreensível para o usuário (TRAINA et al., 2001). Os sistemas de *data mining* procuram integrar a capacidade de exploração do usuário com os recursos computacionais, de forma a produzir um ambiente que propicie a descoberta de conhecimento. A metodologia baseia-se na funcionalidade característica das estruturas internas dos dados e na exibição dos mesmos e, por outro lado, na capacidade do ser humano em perceber padrões, exceções, tendências e relacionamentos ao analisar os dados minerados pelo sistema (KEIM, 2002).

No contexto da saúde, um processo de análise e mineração de dados conforme o descrito tem como potencial prover uma perspectiva global, onde pode-se observar uma distribuição geral dos dados e entender e comparar dados de diferentes estudos. Além disso, sob uma perspectiva local permite a compreensão das relações e melhora a compreensão da evolução dos caminhos clínicos.

As técnicas de mineração de dados em saúde abrangem processos administrativos e de diagnóstico, tratamento e prevenção de doenças, lesões e outras deficiências físicas e mentais em humanos (YANG et al, 2015). A evolução tecnológica no setor de saúde sugere que a mineração pode ser considerada uma fonte primordial de interpretação de informações obtidas em registros médicos eletrônicos e relatórios administrativos (WICKRAMASINGHE; SHARMA; GUPTA, 2008). A mineração de dados, portanto, apresenta-se como uma ferramenta capaz de evidenciar informações novas e valiosas muitas vezes “escondidas” em acervos com grandes volumes de dados. Sua utilidade envolve desde a otimização da prevenção de doenças e o aprimoramento do diagnóstico médico e tomada de decisão clínica até a melhoria da eficiência de práticas administrativas (redução de custos, por exemplo).

Apesar de sua importância e potencial contribuição, a mineração de dados na área da saúde continua sendo, em grande parte, um exercício “apenas” acadêmico com algumas histórias de sucesso pontuais. Conforme destacam Jothi e Husain (2015) e Dhillon e colaboradores (2021) é evidente o interesse acadêmico neste tema, no entanto é evidente que os cuidados com a saúde podem incorporar as pesquisas mais recentes de mineração de dados na prática cotidiana, conforme destaca Haughom (2014).

A mineração de dados possui um grande potencial para a área de saúde, permitindo que os sistemas de saúde usem, de modo sistemático, dados e análises para identificar ineficiências e boas práticas que melhorem os cuidados dos pacientes e reduzam os custos. Alguns especialistas (MARTINEZ; KING; CAUCHI, 2016) acreditam que as oportunidades de melhorar os cuidados e reduzir os custos simultaneamente podem se aplicar a até 30% dos gastos gerais com saúde. Em um cenário ideal, todos os sistemas de saúde deveriam ter os dados históricos necessários para a aplicação de algoritmos de mineração, que permitissem análises preditivas para reduzir gastos e melhorar a qualidade dos atendimentos. Entretanto, no cenário real, as coisas podem ser um pouco mais complexas. Os sistemas de saúde nem sempre têm os dados históricos necessários para aplicação de mineração de dados. O sistema de saúde, de forma geral, precisa melhorar a documentação primeiro e gerar, de forma sistemática e padronizada, os dados necessários antes de iniciar a análise preditiva.

O desempenho dos métodos de mineração de dados varia de acordo com as características e o tamanho dos conjuntos de dados. Observa-se que, de um modo geral, os dados de natureza médica, formam conjuntos desbalanceados o que dificulta as tarefas de mineração de dados (BARELLA, 2021). Outra característica desses conjuntos de dados são os valores ausentes e a quantidade reduzida de amostras. Não existe um método de mineração de dados adequado para resolver todos esses problemas, desta forma é comum a utilização de mais de um algoritmo de mineração para analisar um mesmo acervo de dados. Apesar dessas dificuldades a mineração deve ser amplamente utilizada na análise de dados na área de saúde frente aos potenciais benefícios que ela pode apresentar. Além disso, é importante que a mineração de dados saia da aplicação apenas acadêmica e passe para a parte da prática clínica cotidiana, na busca de um processo de melhoria contínua da

qualidade da saúde pública. Embora esses modelos preditivos exijam uma equipe multifuncional comprometida (profissionais da área de saúde e informática) e precisem ser testados ao longo do tempo, os resultados de pesquisas preliminares reforçam o potencial do uso desta tecnologia (DHILLON et al., 2021).

A estratégia mais eficaz para levar a mineração de dados além do campo da pesquisa acadêmica é a abordagem dos três sistemas (Análise, Melhores Práticas, e Adoção) destacada por Haughom (2014). Entretanto, atualmente, poucas organizações de saúde implementam todos esses três sistemas:

- O sistema de análise inclui a tecnologia para coletar dados (dados clínicos, financeiros, de satisfação do paciente e outros), e a experiência para compreendê-los (interpretá-los) e padronizar as medições;
- O sistema de melhores práticas envolve a padronização do trabalho, aplicando sistematicamente as melhores práticas baseadas em evidências científicas para a prestação de cuidados. Os pesquisadores fazem descobertas significativas a cada ano sobre as melhores práticas, mas, nem sempre são incorporadas à prática clínica. Um forte sistema de melhores práticas permite que as organizações utilizem as mais recentes evidências médicas, utilizando a tecnologia como aliada nesta tarefa;
- O sistema de gerenciamento de mudanças envolve a adoção de novas estruturas organizacionais. Em particular, envolve a implementação de estruturas de equipe que permitirão a adoção consistente em toda a empresa das melhores práticas. Este sistema não é fácil de implementar. Requer uma mudança organizacional real para impulsionar a adoção das melhores práticas em toda a organização.

Se uma iniciativa de mineração de dados não envolver todos esses três sistemas, é provável que continue sendo um exercício puramente acadêmico, sem a implementação prática no dia a dia profissional. A implementação dos três sistemas permite que uma organização de saúde incorpore a mineração de dados à prática clínica diária.

Particularmente na área das anomalias craniofaciais existe um interesse da comunidade científica em identificar estratégias globais para reduzir a

sobrecarga/custo (*burden of care*) do gerenciamento das anomalias (SHAW, 2004). Dentre as estratégias recomendadas pela própria OMS (Organização Mundial da Saúde) destaca-se a necessidade de criação, desenvolvimento, acesso e mineração de dados pré-existentes que possam impulsionar o conhecimento da ciência e a cooperação entre grupos de pesquisa (SHAW, 2004).

Neste sentido busca-se com este trabalho, aplicar a mineração de dados para um melhor entendimento do acervo estabelecido no Estudo Clínico Randomizado (ECR) realizado no Hospital de Reabilitação de Anomalias Craniofaciais da USP (HRAC-USP). O ECR é um estudo conhecido como Projeto Florida (PF), cujo desenvolvimento do acervo de dados teve início nos anos 90 por meio de uma parceria entre o HRAC-USP e o Centro Craniofacial da Universidade da Florida (University of Florida Craniofacial Center, UFCFC, Gainesville, FL, USA). Considerando-se a abordagem similar que faz uso de três sistemas proposta por Haughom (2014), o ECR-PF possibilitou implementar no HRAC/USP tanto a tecnologia quanto os protocolos para coletar dados clínicos e de satisfação do paciente, por exemplo, além de impulsionar o desenvolvimento de habilidades da equipe de pesquisadores para padronizar as medições e interpretar os achados.

De uma forma geral, este trabalho possibilitará ampliar as parcerias entre as áreas de computação e ciências da saúde, sobretudo no escopo das fissuras labiopalatinas, buscando otimizar processos relacionados às práticas dessas ciências, por meio da aplicação de conceitos relacionados à inteligência artificial, ao aprendizado de máquina e à mineração de dados no contexto do ECR-PF no HRAC/USP. Com as tarefas de mineração propostas neste estudo busca-se estabelecer um sistema que possibilite a adoção de melhores práticas, a partir da identificação de evidências científicas sobre a prestação dos cuidados observados ao longo de mais de 20 anos de gerenciamento de grupo de cerca de 466 pacientes, todos com fissura transforame unilateral (FTU) e sem síndromes associada. Ou seja, busca-se amplificar correlações e padrões nos dados que permitam fornecer *insights* ao tratamento da Fissura Labiopalatina (FLP).

2

Referencial
teórico

2 REFERENCIAL TEÓRICO

Neste capítulo é apresentada a fundamentação teórica associada aos conceitos discutidos nesta tese. O foco está na definição da mineração de dados e nos diferentes paradigmas de aprendizagem automática e tarefas de mineração, bem como, conceitos sobre Inteligência Artificial. Também são descritos alguns exemplos sobre a aplicação da mineração de dados na área de saúde.

2.1 INTELIGÊNCIA ARTIFICIAL, APRENDIZAGEM AUTOMÁTICA E MINERAÇÃO DE DADOS

Define-se Inteligência Artificial (IA) como um processo de automação do comportamento inteligente (LUGER, 2004). Segundo Russel e Norvig (2004) a IA é uma área que visa construir agentes racionais para a resolução de problemas. Um agente racional seria toda entidade capaz de perceber o ambiente por meio de sensores e atuar sobre o mesmo, racionalmente, por meio de atuadores. Sobretudo a partir da década de 1970, houve maior disseminação do uso de técnicas de IA para a resolução de problemas reais. Muitos desses problemas eram tratados tendo como base a aquisição de conhecimento de especialistas do domínio. Esse conhecimento era então codificado e disponibilizado na forma de algoritmos que eram base dos chamados Sistemas Especialistas ou Sistemas Baseados em Conhecimento (GAMA *et al.*, 2017). O processo de aquisição de conhecimento era pautado em entrevistas com especialistas com o intuito de descobrir as regras que deveriam ser usadas em processos de tomada de decisão. Entretanto, esse processo apresentava várias limitações associadas aos especialistas tais como subjetividade, dificuldade de comunicação e verbalização do conhecimento, uso da intuição (PRESSMAN, 2006).

Nas últimas décadas com o crescimento da complexidade dos problemas a serem tratados computacionalmente aliado ao enorme volume de dados gerados, houve a necessidade de se desenvolver ferramentas computacionais mais sofisticadas e autônomas que reduzissem a necessidade de intervenção humana e a grande dependência dos especialistas. Neste contexto, novas áreas como a mineração de dados ganharam força e destaque pois são baseadas em técnicas capazes de criar, com base em dados históricos armazenados, uma hipótese ou

função capaz de resolver algum tipo de problema. O processo de descoberta de uma hipótese pode ser representada ao final do processo de mineração, por exemplo, como um conjunto de regras. Segundo Gama e colaboradores (2017), esse processo de indução de uma hipótese (e não da aplicação direta de um algoritmo passo a passo previamente definido) a partir de dados ou experiências passadas corresponde aos objetivos do processo de mineração de dados.

Na Computação é possível resolver muitos problemas por meio da escrita de um algoritmo cuja aplicação, passo a passo, descreve uma resolução (GAMA *et al.*, 2017). Entretanto, para algumas tarefas, mesmo coisas do dia a dia, não é fácil definir algoritmos que sejam capazes de resolver problemas com eficiência. Por exemplo, a tarefa de reconhecimento de pessoas pelo rosto ou pela fala. Quais características devem ser levadas em conta? Como codificar por meio de algoritmos diferentes aspectos como expressões faciais de uma mesma pessoa, ou ainda, alterações na face (por exemplo, bigodes, maquiagens, óculos, cortes de cabelo) ou alterações na voz, devido a uma gripe, uma deformidade ou o humor? Apesar de todas essas dificuldades, os seres humanos realizam essa tarefa com relativa facilidade. Fazem isso por meio do reconhecimento de padrões, uma vez que tenham aprendido o que deve ser considerado no rosto ou na fala para reconhecer uma pessoa, após terem tido vários exemplos de rostos e falas com identificação clara (GAMA *et al.*, 2017).

Outro exemplo da utilização de padrões são os casos de médicos que conseguem diagnosticar problemas de saúde em pacientes, tendo por base um conjunto de sintomas e resultados de exames clínicos. Para este feito, o médico utiliza sua experiência prática e o conhecimento adquirido durante sua formação. Não é simples escrever algoritmos que sejam capazes de atuar de forma similar ao trabalho humano. Muitas vezes existem milhares de dados disponíveis, mas que sem o devido processamento, a escrita de algoritmos que favorecem um melhor entendimento das informações obtidas nem sempre é possível. As técnicas de mineração de dados têm sido utilizadas com sucesso neste sentido (DHILLON *et al.*, 2021).

Conforme destacam Ferrari e Castro (2016), o termo mineração de dados foi cunhado numa alusão ao processo da extração de minerais valiosos a partir de

uma mina. Computacionalmente, refere-se à exploração de uma base de dados, usando algoritmos para obter conhecimento. Esse processo é realizado com o intuito de obter conhecimento útil, implícito e previamente desconhecido sobre os dados armazenados (WITTEN; FRANK; HALL, 2016). A ideia é utilizar, para esta atividade, programas de computador ou linguagens de programação que processem automaticamente (ou de forma semiautomática) as bases de dados, buscando por regularidades ou padrões. Bons padrões, se encontrados, possibilitam fazer predições acuradas frente a dados futuros (WITTEN; FRANK; HALL, 2016).

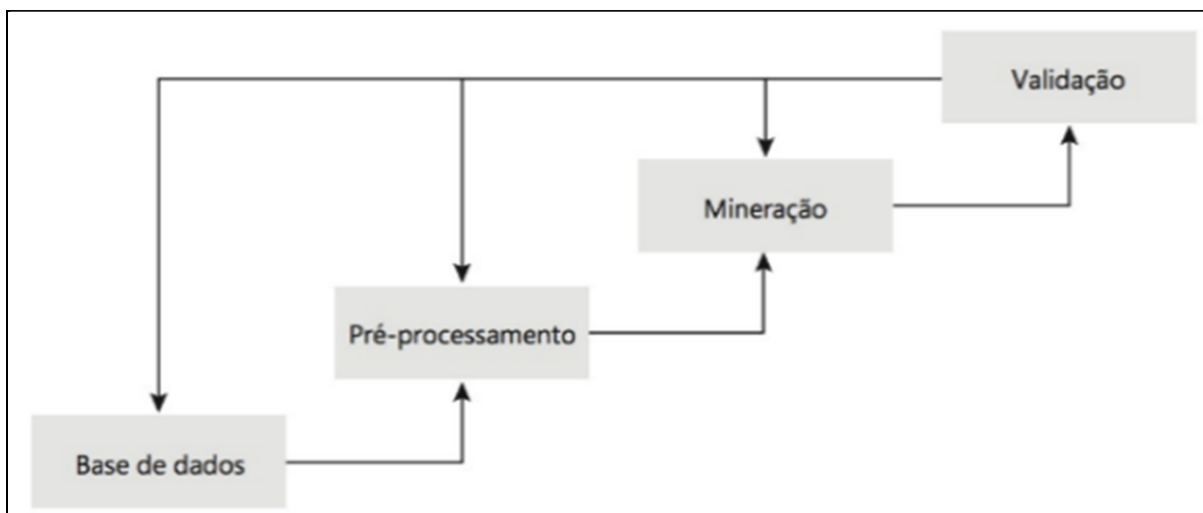
Buscar padrões é uma atividade humana e usada historicamente. Caçadores buscam padrões no comportamento da migração de animais, agricultores buscam padrões no crescimento de lavouras, políticos buscam padrões na opinião dos eleitores. Profissionais da saúde buscam padrões no gerenciamento das condições que afetam o bem-estar. Buscar padrões permite às pessoas dar sentido aos dados e entender melhor como o mundo funciona, encapsulando esse conhecimento em teorias que poderão ser utilizadas para prever o que irá ocorrer em situações futuras. Esses padrões são expressos de duas formas: como “caixas-pretas” cujas entranhas são praticamente incompreensíveis ou como uma “caixa-transparente” cuja construção revela a estrutura do padrão. Ambas as formas, podem ser usadas para boas previsões. A diferença é se os padrões que são minerados são ou não representados em termos de uma estrutura que pode ser examinada, analisada e usada para informar futuras decisões. Tais padrões são chamados de padrões estruturados porque eles capturam a estrutura da decisão de uma maneira explícita. Em outras palavras, os padrões estruturados ajudam a explicar algo sobre os dados. Por exemplo, identificar padrões de resultados do tratamento da fissura labiopalatina pode otimizar o gerenciamento desta condição ao permitir a identificação de fatores que podem predizer aspectos da fala, da função velofaríngea, do crescimento e da estética facial após a palatoplastia primária, por exemplo. Neste sentido, a busca por variáveis que carregam grande acurácia preditiva é um valioso recurso para otimizar a tomada de decisão.

Cabe destacar que a mineração de dados é parte integrante de um processo mais amplo, definido como KDD (*Knowledge Discovery in Databases*) –

descoberta de conhecimento em bases de dados (FERRARI e CASTRO, 2016). De acordo com o que foi definido na primeira conferência internacional sobre KDD, realizada em 1995 na cidade de Montreal, Canadá, a terminologia descoberta de conhecimento em bases de dados refere-se a todo o processo de extração de conhecimento a partir dos dados. O termo mineração de dados deve ser empregado exclusivamente para a etapa de descoberta do processo de KDD. O KDD, que por ser mais amplo, inclui outras atividades além da etapa de descoberta, incluindo a seleção e integração das bases de dados, a limpeza dos dados, a seleção e transformação dos dados e, por fim, a mineração e avaliação dos dados.

Como propõem Ferrari e Castro (2016), o processo de KDD pode ser sintetizado em quatro etapas principais, conforme mostra a Figura 1.

Figura 1 – Etapas do processo de KDD



Fonte: Ferrari e Castro (2016, pag. 6).

Conforme pode ser visto na Figura 1, o processo se inicia com a seleção de uma base de dados que pode ser vista como o insumo principal deste processo. A base de dados corresponde a uma coleção organizada de dados, isto é, valores quantitativos ou qualitativos referentes a uma amostra de um conjunto de itens. Conceitualmente os dados correspondem ao nível mais básico da abstração a partir do qual informação e, posteriormente, conhecimento, podem ser obtidos. A etapa seguinte deste processo corresponde ao pré-processamento (preparação) dos dados. Esta etapa visa preparar os dados para que a análise se dê de forma

eficiente e eficaz. Inclui a remoção de ruídos ou dados inconsistentes (tarefa chamada de limpeza), a integração (combinação) de dados obtidos de múltiplas fontes, a seleção ou redução que corresponde a uma escolha de dados que sejam relevantes à análise) e a transformação ou consolidação dos dados em formatos apropriados para a mineração).

A etapa de mineração corresponde à aplicação de algoritmos capazes de extrair conhecimento. Envolve também a descrição dos dados (análise descritiva) por meio de medidas de distribuição, tendência central e variância, por exemplo; agrupamento (segmentação ou divisão em conjuntos); predição (classificação e estimação), associação (verificação de atributos que coocorrem) e detecção de anomalias (busca de *outliers*, isto é, dados que se diferenciam drasticamente de todos os outros). A última etapa do processo de KDD refere-se à validação (ou avaliação) do conhecimento, que visa determinar se o mesmo é verdadeiramente útil e não trivial. Todas as etapas do KDD são inter-relacionadas e influenciam diretamente o processo de extração de informações relevantes em bases de dados.

Outro destaque feito por Ferrari e Castro (2016), refere-se à caracterização da mineração de dados como uma disciplina interdisciplinar e multidisciplinar já que envolve conhecimento de áreas diversas como estatística, banco de dados, aprendizagem de máquina, inteligência artificial, probabilidade, dentre outras (WITTEN; FRANK; HALL, 2016; BHATTACHARYYA e HAZARIKA, 2006; KAREGAR *et al.*, 2008; HILL e LEWICKI, 2005). Neste sentido, o aprendizado de máquina (em inglês, *machine learning*) é um subcampo da Ciência da Computação relacionado ao reconhecimento de padrões que visa dar aos computadores a habilidade de aprender sem serem explicitamente programados. Para isso, os algoritmos utilizados usam dados amostrais (previamente conhecidos e rotulados) para fazer previsões sobre novos dados. Mineração de dados e aprendizado de máquina, apesar de serem conceitos bastante similares, têm focos distintos. A mineração de dados é projetada para extrair as regras de grandes quantidades de dados, enquanto o aprendizado de máquina visa “ensinar” um computador a aprender e compreender os parâmetros fornecidos. Ou seja, a mineração de dados é simplesmente um método de pesquisa para determinar um

resultado específico com base no total dos dados coletados. Por outro lado, o aprendizado de máquina treina um sistema para realizar tarefas complexas e usa os dados e a experiência coletados para se tornar mais inteligente. A mineração de dados tem na intervenção humana algo muito mais direto e, em última análise, seus resultados são destinados para serem usados por pessoas (por exemplo, um padrão descoberto pode ser usado por um profissional de saúde como apoio ao seu processo de tomada de decisão). Já o aprendizado de máquina visa certa independência do processo, podendo “ensinar a si mesmo” e depender minimamente da influência ou ações humanas. O aprendizado de máquina está diretamente ligado à área de Inteligência Artificial (IA), probabilidade e estatística. A IA é responsável por estudar teorias que permitam simular comportamentos inteligentes nas máquinas. A probabilidade envolve o estudo matemático na quantificação da aleatoriedade e incerteza de eventos, já a Estatística é a ciência da coleta, descrição e análise de dados. O foco desta tese está nas técnicas de mineração de dados, visando um processo de indução de modelos.

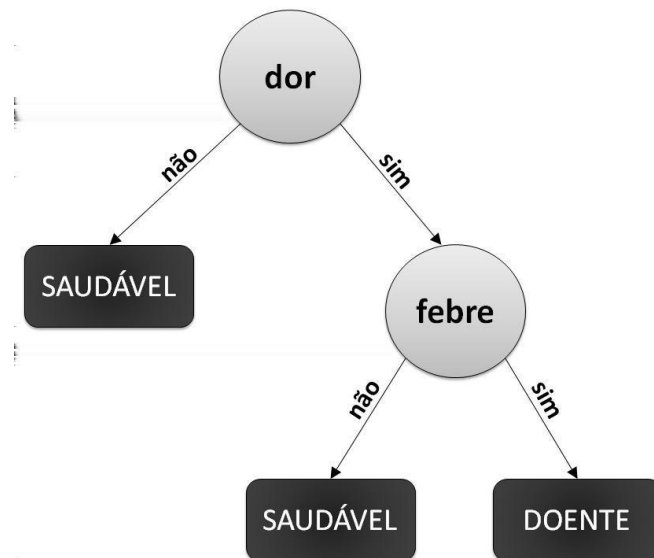
Neste sentido os algoritmos utilizados empregam um princípio de inferência denominado indução, no qual se obtêm conclusões genéricas a partir de dados amostrados do problema a ser resolvido (WITTEN; FRANK; HALL, 2016). Assim, os modelos gerados são capazes de lidar com situações não apresentadas durante o seu desenvolvimento. Várias áreas, incluindo-se a saúde, podem se beneficiar das características da área de mineração de dados e da elevada capacidade dos recursos computacionais atualmente disponíveis (softwares e hardware).

Na mineração de dados, mais especificamente, a indução de hipóteses se dá a partir de um conjunto de dados observados, como por exemplo, dados sobre pacientes de uma clínica craniofacial. Cada paciente é denominado objeto e sobre cada objeto são armazenados diversos atributos (nome, identificação, sexo, idade, sintomas etc.), que correspondem às diversas características desse paciente. Em uma das tarefas típicas da mineração busca-se aprender formas de se prever um dos atributos (esse atributo específico que se deseja fazer a predição é denominado de classe ou, simplesmente, atributo alvo ou atributo de saída). Os demais atributos utilizados para se fazer a predição do atributo alvo são chamados de preditores ou

atributos de entrada. A partir de um conjunto de dados busca-se criar um modelo ou hipótese (representados por um algoritmo ou conjunto de regras) capaz de relacionar um ou mais atributos (preditores) ao atributo alvo (classe).

Por exemplo, na Figura 2 um modelo é representado por meio de uma árvore de decisão (gerada automaticamente por um algoritmo de mineração de dados), cuja interpretação permite extrair regras que representam o modelo de relacionamento entre os atributos. A “leitura” da árvore permite concluir que se o paciente tiver dor (dor=sim) e febre (febre=sim) a conclusão é que ele está doente. Por outro lado, se não houver dor (dor=não) ou ainda se houver dor (dor=sim) mas não houver febre (febre=não) o mesmo poderia ser considerado saudável. No exemplo, os atributos de entrada (preditores) seriam “dor” e “febre” e o atributo alvo (classe ou atributo de saída) seria a “situação do paciente” (saudável ou doente). Cada atributo é chamado de nó da árvore. Os atributos de entrada são referidos como nós internos e o atributo de saída é chamado de nó externo (folha).

Figura 2 – Exemplo de árvore de decisão



Fonte: Elaborada pelo autor

Por exemplo, modelos como este, em forma de árvores de decisão (geradas automaticamente), utilizam uma estrutura de árvore para associar a cada um de seus nós internos (círculos) uma pergunta referente aos valores dos atributos

e cada nó externo (retângulo) está associado à uma classe (saída). As setas representam os valores que cada nó (atributo) pode assumir, no exemplo ilustrado na Figura 2, o atributo “dor” pode assumir os valores “sim” ou “não”.

Por meio de um viés indutivo, cada algoritmo identificado a partir da mineração de dados utiliza uma representação para descrever a hipótese induzida a partir do conjunto de dados. O conjunto de dados corresponde a toda a amostra da base de dados que será utilizada no processo de mineração. Este conjunto de dados pode ser dividido randomicamente em dois outros conjuntos: conjunto de treinamento e conjunto de testes. O conjunto de treinamento é usado para construir o modelo e o conjunto de testes é usado para verificar a qualidade do modelo gerado (estimar o seu erro) por meio de alguma métrica. A separação do conjunto de dados em treinamento e teste permite verificar a capacidade de generalização do modelo já que os testes serão realizados considerando dados que o modelo não teve contato durante o seu processo de criação. No exemplo dado, o objetivo da mineração poderia ser induzir uma hipótese capaz de realizar diagnósticos corretos para novos pacientes, considerando a capacidade de generalização do modelo. Uma boa hipótese é aquela com uma boa capacidade de generalização. Modelos pouco genéricos superajustados aos dados (*overfitting*) ou subajustados (*underfitting*) não são desejados e são frutos de uma provável pouca representatividade dos dados de treinamento, sendo incapazes de capturar padrões (MONARD e BARANAUSKAS, 2003 *apud* GAMA *et al.*, 2017). Numa analogia aos seres humanos quando estamos aprendendo a ler letras cursivas temos contato apenas com algumas caligrafias (conjunto de treinamento), entretanto, uma vez aprendido o processo de leitura reconheceremos a caligrafia de qualquer pessoa (conjunto de testes), mesmo que nunca tenhamos visto a letra daquela pessoa antes. Isso ocorre devido a nossa capacidade de generalização e é essa característica que buscamos nos modelos de mineração de dados.

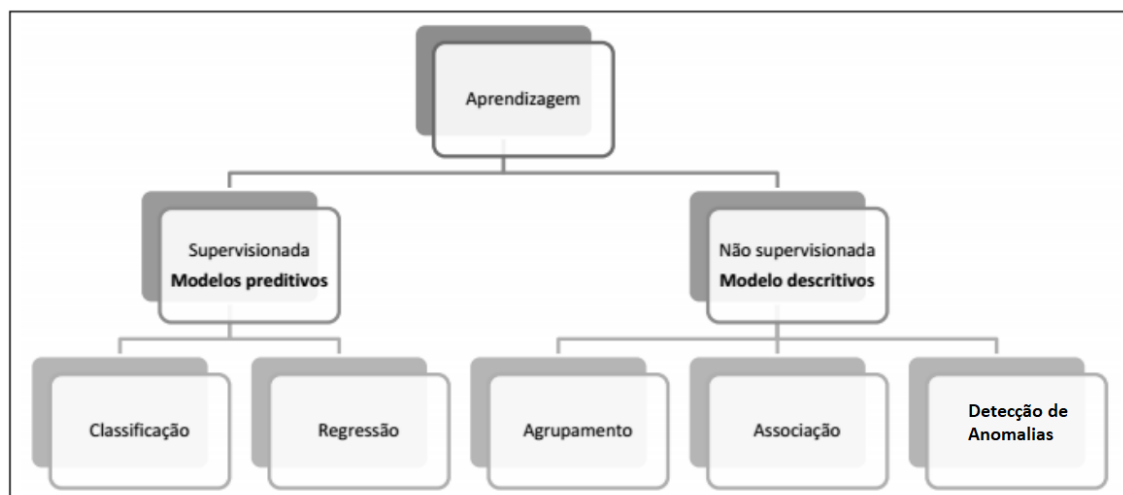
Além do viés de representação (na forma de árvore de decisão, por exemplo), cada algoritmo apresenta diferentes vieses de procura, isto é, a forma como o modelo busca a hipótese que melhor se ajusta aos dados de treinamento. No caso de uma árvore de decisão este viés de procura corresponde a utilizar diferentes algoritmos para escolher como a árvore será formada (em quais posições

estarão cada um dos atributos na árvore). Conforme Mitchell (1997 apud GAMA et al., 2017, p. 14) esses vieses de representação e procura é o que permite cada algoritmo generalizar o conhecimento adquirido.

2.2 PARADIGMAS DE APRENDIZAGEM E TAREFAS DA MINERAÇÃO

No contexto da mineração de dados define-se como aprendizagem o processo pelo qual os computadores desenvolvem o reconhecimento de padrões ou a capacidade de aprender continuamente com os dados ou fazer previsões neles baseadas. Considerando os diferentes paradigmas (formas) de aprendizagem, as tarefas podem ser classificadas em tarefas de aprendizagem supervisionadas ou não supervisionadas. Outra classificação envolve ainda diferenciar as tarefas de mineração de dados de acordo com o tipo de informação a ser obtida incluindo: modelos de preditivos (classificação e regressão) e modelos descritivos (agrupamento, associação e detecção de anomalias) (FERRARI e CASTRO, 2016). Essas formas de classificação e suas associações podem ser vistas na Figura 3.

Figura 3 – Classificação de tarefas de aprendizagem



Fonte: Elaborada pelo autor

Nos modelos preditivos o objetivo consiste em encontrar, a partir dos dados, uma hipótese (modelo ou função) que possa ser utilizada para prever a classe ou valores de novos exemplos não vistos durante a fase de treinamento do algoritmo. Isto é, deseja-se criar um modelo preditor capaz de classificar um novo objeto com base nos seus atributos. Para que isso seja possível é necessário que cada objeto do conjunto de dados tenha atributos de entrada e um atributo alvo (classe) já definido. O modelo preditivo geralmente aplica funções de aprendizagem supervisionada para prever valores desconhecidos ou futuros de outras variáveis de interesse (KANTARDZIC, 2019).

O termo aprendizagem supervisionada faz uma alusão à presença de um "professor/supervisor" que fornece dados previamente rotulados, indicando a que classe aqueles dados pertencem. Na prática imagine que se deseja classificar pacientes em saudáveis e não-saudáveis e, para fazer isso, é disponibilizada uma amostra que associa cada paciente saudável a uma série de variáveis (atributos). Assim, um algoritmo de aprendizagem supervisionado tentaria usar explicitamente essa informação para, no futuro, ser capaz de separar pacientes saudáveis de não-saudáveis. Neste caso, o supervisor é capaz de avaliar, com base neste atributo alvo, o desempenho do algoritmo, indicando, por exemplo, a taxa de acertos, isto é, quantos objetos foram corretamente classificados (por exemplo, dos pacientes indicados como saudáveis, quantos, de fato, são saudáveis?).

O modelo preditivo visa a construção de um modelo que permita tanto classificar um objeto novo ainda não rotulado (classificação) quanto inferir o valor numérico de um ou mais atributos do objeto em questão (regressão). Sob essa ótica, a classificação é usada para prever valores discretos enquanto a regressão é usada para prever valores contínuos. Como exemplo considere o problema de diagnosticar uma doença, no qual um paciente se dirige a um hospital para saber se está doente e se deve tomar algum remédio. A primeira pergunta a ser respondida está associada a uma tarefa de classificação: o paciente está ou não doente? A segunda pergunta seria: uma vez doente, qual o medicamento e a dose deste que deve ser ministrada? Esta última corresponde a uma estimação (tarefa de regressão) que faz sentido na medida em que o paciente se encontra doente e precisa ser medicado. Como os rótulos (atributos) dos dados de treinamento são

previamente conhecidos, eles são usados para induzir um modelo de predição capaz de dizer se um paciente está ou não doente. Esse processo se dá seguindo a aprendizagem supervisionada. Os modelos preditivos são mais comumente usados na área da saúde (JOTHI e HUSAIN, 2015).

Nos modelos descritivos o objetivo é explorar ou descrever o conjunto de dados. Nos algoritmos utilizados neste tipo de tarefa o atributo de saída (atributo alvo) é ignorado. Deste modo, este modelo segue a denominada aprendizagem não supervisionada. Neste tipo de aprendizagem não existe um rótulo de saída desejado, ou seja, não são fornecidos dados já previamente rotulados (GAMA *et al.*, 2017; KANTARDZIC, 2019). No modelo descritivo, portanto, o objetivo é sumarizar os dados, isto é, compreender os objetos da base de treinamento e seus atributos. O foco é medir, explorar e descrever características intrínsecas aos dados. Essa análise permite investigar medidas de centro e variação, distribuição de frequências, média, desvio-padrão, moda, medidas de posição relativa e associação dos dados. Além disso, técnicas de visualização também estão associadas a esta tarefa de descrição. As informações da descrição podem ser representadas por meio de gráficos, explicitando um conhecimento específico sobre os dados (WITTEN; FRANK; HALL, 2016).

Na tarefa de análise de grupos (ou agrupamento, do inglês *clustering*) o objetivo é segmentar (distribuir) o conjunto de dados em grupos (clusters) de objetos similares (FERRARI e CASTRO, 2016). Esta tarefa considera que os dados de treinamento não estão previamente rotulados, ou seja, as classes (grupos) aos quais cada objeto pertence não são conhecidas, a priori. O processo de agrupamento é utilizado para a identificação dessas classes e cada grupo formado indica, portanto, uma classe definida pelo algoritmo. Por não se conhecer de antemão a classe de cada objeto, esse processo se enquadra no paradigma do aprendizado não supervisionado. No agrupamento visa-se minimizar a distância intraclasse (dentro do grupo) e, ao mesmo tempo, maximizar a distância interclasse (entre objetos de diferentes grupos), garantindo que os grupos estejam bem definidos. Para ilustrar uma tarefa de agrupamento, considere um problema em que se deseja segmentar (distribuir) uma base de pacientes hospitalizados, na qual cada paciente está

descrito por um conjunto de atributos. Suponha que haja pacientes em leitos comuns e outros em UTI e que o algoritmo precisa distribuí-los nos diferentes tipos de tratamento sem ter conhecimento algum sobre quais tipos de leitos eles ocupam, recebendo apenas os demais atributos dos pacientes. Como alguns atributos associados a pacientes em UTI se diferem daqueles que estão em leitos comuns, durante o agrupamento, o algoritmo provavelmente irá colocá-los em grupos distintos. Veloso e colaboradores (2014), por exemplo, utilizaram o método de quantização vetorial na abordagem de agrupamento para prever os retornos de pacientes na medicina intensiva. Os conjuntos de dados usados neste estudo foram coletados da avaliação clínica dos pacientes e de exames laboratoriais. A partir dos resultados, o trabalho desses pesquisadores forneceu diretrizes úteis para ajudar a identificar os pacientes com maior probabilidade de serem readmitidos.

Na tarefa de associação o objetivo é encontrar relações entre atributos e não entre objetos. Neste caso, busca-se minerar regras de associação que indicarão valores de atributos que ocorrem concomitantemente em uma base de dados (FERRARI e CASTRO, 2016). Por exemplo, pode-se descobrir via tarefa de associação que dois sintomas costumam ocorrer em casos de uma certa doença (quem tem febre, também tem dor no corpo, por exemplo). Na mineração de regras de associação os algoritmos visam propor regras que sejam significativas de modo eficiente. Esta significância das regras está associada à capacidade de detectar associações que sejam estatisticamente relevantes para o universo da base de dados.

Nas tarefas de detecção de exceções (“anomalias”) os algoritmos procuram detectar nas bases de dados objetos que não seguem um comportamento ou não possuem uma característica comum dos dados ou de um modelo que os represente. Na Computação o termo anomalias se refere aos *outliers*. Os *outliers* são dados que se diferenciam drasticamente de todos os outros, são pontos “fora da curva normal”. Em outras palavras, um *outlier* é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise. Entender os *outliers* é fundamental em uma análise de dados por pelo menos dois aspectos: os *outliers* podem enviesar negativamente todo o resultado de uma análise; o comportamento dos outliers pode

ser justamente o que está sendo procurado. Geralmente as tarefas de mineração costumam descartar (identificar) as anomalias (também definidas como outliers, ruídos, exceções, algo que foge da regra) e, em algumas situações, como na detecção de doenças raras, por exemplo, os eventos ou características diferentes podem ser mais informativos do que aqueles que ocorrem de forma regular. Para a detecção de anomalias normalmente são usados métodos estatísticos relacionados à distribuição dos dados ou modelos baseados em medidas de distância, em que objetos considerados muito distantes dos demais são tratados como anomalias (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996; LIU *et al.*, 2013). As anomalias compõem uma classe que ocorre com uma frequência notadamente inferior às demais. Desta forma, os algoritmos de classificação são fortemente impactados pelas anomalias, forçando o uso de algoritmos e medidas de desempenho desenvolvidos especificamente para lidar com esses casos em que existe um desbalanceamento de dados, isto é, os dados de interesse ocorrem numa proporção bastante inferior aos demais na base de dados.

2.3 A MINERAÇÃO DE DADOS NA ÁREA DE SAÚDE

Várias técnicas de mineração de dados foram desenvolvidas ao longo dos últimos anos e vêm ganhando cada vez mais espaço com aplicações na área de saúde (LIAO; CHU e HSIAO, 2012; DHILLON *et al.*, 2021). Conforme revisão sistemática desenvolvida por Jothi e Husain (2015) e Dhillon e colaboradores (2021) podem ser encontrados vários trabalhos científicos que corroboram com essa percepção de disseminação de técnicas de mineração de dados na área de saúde, em que a tecnologia tem um papel importante, especialmente no desenvolvimento de metodologias para a coleta e análise de dados. Uma vez definida a técnica de mineração de dados, são escolhidos os algoritmos a serem utilizados de acordo com a adequação ao problema a ser resolvido. Para a tarefa de classificação, por exemplo, alguns algoritmos amplamente utilizados incluem: árvores de decisão, K-vizinhos mais próximos (K-NN), MLP (Multilayer Perceptron, um tipo de rede neural artificial), vetores de suporte e redes bayesianas (WITTEN; FRANK; HALL, 2016).

Sistemas de mineração de dados podem ser utilizados para auxílio às funções administrativas ligadas à área de saúde. Por exemplo, modelos de mineração podem ser usados para definir categorias de planos de saúde, mantendo, ao mesmo tempo, seu volume de pacientes estável e definindo os melhores processos para garantir que esses pacientes recebam os cuidados apropriados no lugar e hora certos. Isso incluiria gerenciamento de cuidados para pacientes de alto risco. Outra possibilidade é a utilização desses sistemas para prever quais pacientes deverão retornar às unidades hospitalares após terem recebido alta, de modo a permitir o acompanhamento adequado. Outras iniciativas envolvem a aplicação de algoritmos preditivos aos dados de bases para prever riscos em determinadas populações. Usando os dados, portanto, é possível identificar os parâmetros clínicos e demográficos com maior probabilidade de prever um evento de atendimento para uma população específica. Esse processo de estratificação de pacientes em grupos de alto, médio ou baixo risco é essencial para o sucesso de qualquer iniciativa de gestão da saúde da população. Em certas situações, alguns pacientes correm tanto risco que seria mais barato enviar preventivamente um médico para fazer atendimento em casa, em vez de esperar que esse paciente chegue às unidades hospitalares já em situação crítica. Deste modo, é possível identificar esses pacientes de alto risco com antecedência e concentrar os recursos apropriados em seus cuidados. Isso permite o desenvolvimento de processos aprimorados para gerenciar o atendimento de pacientes em risco. Por exemplo, permitindo que médicos discutam o nível de risco de cada paciente, criando um plano de gerenciamento de cuidados com antecedência para compartilhar com o paciente durante as consultas (JOTHI e HUSAIN, 2015).

Com o uso da mineração é possível, por exemplo, reduzir as taxas de retorno (readmissão) de pacientes em curto prazo, já que esta é outra questão importante que os sistemas de saúde estão enfrentando atualmente. Os algoritmos podem ser usados para criar modelos que identificam os pacientes com altos riscos de retorno (JOTHI e HUSAIN, 2015). Quando o sistema de saúde possui um conjunto de dados históricos adequados (dados sobre pacientes com certas condições recentemente atendidos), a mineração possibilita identificar esses dados para criar um modelo preditivo.

A mineração de dados também pode ajudar o sistema de saúde a otimizar seus esforços, avaliando a eficácia relativa das melhores práticas. Por exemplo, se um profissional tiver apenas tempo para aplicar algumas das intervenções a um paciente, qual intervenção ou combinação de intervenções terá maior impacto? Os testes de laboratório são uma ferramenta de auxílio para que um profissional de saúde decida como tratar um paciente? A mineração de dados pode ajudar o profissional a descobrir informações que de outra forma poderiam ser difíceis de se perceber nos resultados do laboratório. El-Halees e Almadhoun (2017) analisaram mais de 600 amostras de urina e usaram a mineração de dados para classificar os pacientes pela expectativa de vida com base nas características de sua urina. Essa abordagem revelou casos em que os pacientes estão mais doentes do que parecem, permitindo que os médicos tomassem medidas interventivas de forma imediata. Dados sobre reações adversas a medicamentos, por exemplo, podem ajudar médicos a avaliar se a prescrição de um novo medicamento para um paciente otimiza o tratamento da pessoa contribuindo para interromper os efeitos colaterais perigosos.

Uma pesquisa realizada por Hansen e colaboradores (2016) mostrou que a mineração de dados pode ajudar os cientistas a descobrir interações comuns e menos prevalentes entre diferentes drogas. Os pesquisadores analisaram 200 grupos de medicamentos cardiovasculares em mais de 13.500 pacientes durante o estudo. Eles descobriram 87 possíveis interações medicamentosas e, em um grupo específico de medicamentos localizaram sete com efeitos colaterais perigosos. Esta investigação foi para medicamentos cardiovasculares, mas também tem valor para outros produtos farmacêuticos.

Quando um médico prescreve um medicamento, ou um farmacêutico orienta sobre a forma de uso do mesmo, a adesão ao tratamento não é necessariamente garantida. Isto é, a recomendação médica e a orientação farmacêutica não implicam necessariamente que um paciente tome o medicamento conforme as instruções. Kumamaru e colaboradores (2018) utilizaram técnicas de mineração de dados para verificar se algumas informações relacionadas às características demográficas, comorbidades e adesão anterior à medicamentos

mostrariam conexões com a adesão aos medicamentos prescritos. O estudo envolveu uma base de dados de 93.777 pessoas que tomavam estatinas e descobriu várias características relacionadas à probabilidade de adesão a medicamentos a longo prazo. Buscou-se, portanto, associações entre as variáveis e a aderência ideal às estatinas nos 12 meses subsequentes. Utilizando técnicas de regressão logística foram desenvolvidos modelos capazes de identificar quais as variáveis seriam as melhores preditoras de que os pacientes teriam alta adesão às estatinas. Uma forte associação foi observada entre a adesão anterior a medicamentos usados cronicamente e a adesão futura ao uso de estatinas. Os autores concluíram que a adesão anterior a medicamentos crônicos foi um forte preditor de adesão futura a estatinas recém-iniciadas.

Os administradores hospitalares procuram continuamente maneiras de aumentar o desempenho, cortar custos e aumentar a eficiência das equipes que coordenam. Muitos deles recorrem à mineração de dados para atingir esses objetivos, geralmente dependendo de consultores de negócios para aprimorar as práticas atuais por meio de insights orientados por dados. A segurança do paciente e os resultados positivos são, sem dúvida, dois fatores que os administradores hospitalares se preocupam em analisar por meio de mineração. Guilliams (2008) usou um conjunto de dados de registros de alta hospitalar na Bélgica e observou que detalhes, como a duração de uma estada e os tipos de tratamentos recebidos, por exemplo, poderiam prever fatores de risco resultando em informações que permitiram que as equipes de saúde otimizassem a segurança do paciente e reduzissem as taxas de readmissão.

Técnicas de mineração de dados também podem ser utilizadas para reduzir os casos de fraude no seguro de saúde. Um estudo que aplicou a mineração de dados para analisar fraude identificou que as seguradoras poderiam selecionar determinados documentos para uma inspeção mais detalhada e, potencialmente, prevenir fraudes (JOUAKI et al, 2015).

Um estudo desenvolvido por Rémy, Martial e Clémentin (2018) mostrou que a mineração de dados poderia prever a capacidade de um médico em diagnosticar pacientes. Mais especificamente, poderia classificar os médicos especialistas com base na probabilidade de diagnosticar corretamente um problema,

reduzindo assim as taxas de erro. De uma forma geral, os pesquisadores concluíram que a informação proporcionada pela mineração de dados é benéfica para ajudar a formar uma equipe de especialistas para fornecer um diagnóstico multidisciplinar, especialmente quando um paciente mostra sintomas de problemas de saúde específicos e de difícil diagnóstico.

Os pesquisadores Su e colaboradores (2012), usaram o algoritmo do Sistema Mahalanobis Taguchi para projetar um modelo de previsão para úlceras. O algoritmo MTS (Multivariate Time Series) pode ser aplicado em análises estatísticas multivariáveis. A distância de Mahalanobis é usada para construção de modelos estatísticos que permitem distinguir grupos e o espaço de Mahalanobis é usado para representar o grau de anormalidade das observações frente a um grupo de referência conhecido. O teste usado pelos pesquisadores usando esse algoritmo foi realizado em quatro fases, com conjuntos de dados desbalanceados. Dados desbalanceados podem ser definidos pela pequena incidência de uma categoria dentro de um conjunto de dados (classe minoritária) em comparação com as demais categorias (classes majoritárias) (WITTEN; FRANK; HALL, 2016). Na maioria dos casos, isso faz com que se tenha muitas informações a respeito das categorias mais incidentes, e menos das minoritárias, o que pode, em muitos casos, interferir na análise de dados. Entretanto, o desbalanceamento nos dados mostra-se presente em diversas situações e campos do conhecimento, não sendo incomum serem encontrados regularmente e em contextos variados, como em dados relacionados ao diagnóstico de doenças. Os resultados obtidos no trabalho (SU *et al.*, 2012) mostraram que a escala de medição desse algoritmo apresentou um bom desempenho mesmo considerando-se a enorme diferença entre os exemplos normais e anormais. Ou seja, esse algoritmo se mostrou adequado em termos de medida de sensibilidade (proporção de verdadeiros positivos) de dados desbalanceados.

Armañanzas e colaboradores (2013) e Jen e colaboradores (2012) usaram a análise discriminante linear em seus trabalhos. A análise discriminante linear (LDA) é usada para decidir à qual de dois grupos pertencem um determinado acervo de dados estudados (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996),

sendo que o desempenho desse algoritmo é melhor quando os dados estudados estão numa relação linear. No estudo de Jen e colaboradores (2012) o algoritmo da análise discriminante linear permitiu prever a gravidade dos sintomas motores em 69% dos pacientes com doença de Parkinson usando escores de provas clínicas não-motoras. Armañanzas e colaboradores (2013), por sua vez, usaram o algoritmo para relacionar fatores de risco para a identificação precoce de doenças crônicas.

Um dos algoritmos mais utilizados para analisar dados clínicos é o de árvores de decisão. Este algoritmo permite examinar dados e criar a árvore de regras que são usadas para fazer uma previsão e melhorar o desempenho de previsões, em termos de precisão. As árvores são bastante legíveis e ajudam os profissionais a entenderem a natureza das decisões (SHARMA e OM, 2013; WANG *et al.*, 2013; ZOLBANIN *et al.*, 2015). Outro tipo de algoritmo usado para tarefa de classificação é o algoritmo K-NN (BAGUI *et al.*, 2003; ŞAHAN *et al.*, 2007; GARCÍA-LAENCINA *et al.*, 2015). Este algoritmo é um método classificador baseado em comparação entre instâncias de dados em acervos não muito grandes uma vez que o processo de classificação é mais longo do que outros algoritmos. Uma instância é a informação coletada em uma base de dados em algum momento específico referindo-se, portanto, a um exemplar da amostra. De uma forma geral a precisão da classificação dos dados é o elemento mais importante em vez do tempo de classificação, já que a precisão da classificação é necessária para o diagnóstico médico adequado.

Alguns pesquisadores adotaram a regressão logística (RL) em seus estudos (SAMANTA *et al.*, 2009; THOMPSON *et al.*, 2014; MAMIYA *et al.*, 2015). A RL é um método que calcula uma combinação linear entre variáveis de entrada através de uma função logística (GENKIN; LEWIS; MADIGAN, 2007). Como neste modelo de mineração de dados (RL) é necessário um conjunto de dados maior, os resultados reportados pelos pesquisadores citados não foram muito significativos. Para lidar com dados ausentes é possível usar os classificadores Bayesianos que envolvem um algoritmo relativamente eficiente computacionalmente e com boa capacidade de precisão de previsão em grupos de dados menores (WITTEN; FRANK; HALL, 2016).

Finalmente, uma máquina de vetores de suporte (do inglês, *support vector machine-SVM*) é um modelo que apresenta um bom desempenho em tarefas de classificação, tendo como característica a capacidade de minimizar o limite superior do erro de generalização, podendo ser utilizado com eficácia em tarefas de diagnóstico médico conforme reportado nos trabalhos de Fei (2010), Zheng e colaboradores (2014), Kang e colaboradores (2015).

3

Objetivo

3 OBJETIVOS

- Explorar, avaliar e evidenciar as variáveis (*features*) associadas aos pacientes com fissuras transforame unilateral (FTU) para predição da ocorrência de fístulas.
- Usar e implementar métodos clássicos de aprendizado de máquina para estudar e avaliar como as tarefas de mineração de dados de classificação e de associação contribuem na evidenciação de correlação e padrões na base de dados do Estudo Clínico Randomizado – Projeto Florida (ECR-PF).
- Comparar o desempenho de algoritmos de aprendizado de máquina (em termos de acurácia, precisão, cobertura e f-measure) para classificação de fístulas nos participantes no ECR-PF, usando especificamente: Árvore de decisão (decision Tree), K-Vizinhos Mais Próximos (k-Nearest Neighbor – k-NN), Regressão Logística (logistic Regresion), Naive Bayes Gaussiano (Gaussian Naive Bayes), Máquinas de Vetores-suporte (Suport Vector Machine) e Floresta Aleatória (Random Forest)).

4

Relação entre os artigos

4 RELAÇÃO ENTRE OS ARTIGOS

Os artigos apresentados nesta tese apresentam estudos que se complementam uma vez que foram realizados considerando a mesma fonte de dados, sob diferentes perspectivas (descoberta de conhecimento e validação de algoritmos). Neste estudo exploratório as fontes de dados que foram acessadas são referentes ao ECR-PF e aprovação para acesso aos dados foi obtida em 2016 (CEP 1.753.467). Os dados do ECR-PF foram armazenados, desde 1996, no servidor em padrão proprietário, com programa funcional no sistema operacional Windows com exportação feita em formato XLS ou em padrão CSV (comma-separated value) com inserção em planilha eletrônica. Os dados de interesse para este estudo incluíram dados sobre a ocorrência de fístulas que foram obtidos para o total de 466 pacientes com FTU. Os dados anonimizados estudados foram obtidos de pacientes que foram randomizados (usando-se um programa computacional) para receber diferentes protocolos de tratamento cirúrgico incluindo: 1) queiloplastia primária entre 3 e 6 meses de idade com a técnica de Millard (M) ou Spina (S); 2) palatoplastia precoce (9 a 12 meses) ou tardia (>12 meses); 3) palatoplastia primária com a técnica de von Langenbeck (VL) ou de Furlow (F); e 4) para um de quatro possíveis cirurgias (C1, C2, C3, C4).

Uma vez que o acervo de interesse foi identificado no servidor, foram feitas simulações com os dados, aplicando-se algoritmos de mineração de dados, usando-se como ferramenta para extração de conhecimento a WEKA (WITTEN; FRANK; HALL, 2016) e a linguagem de programação Python (CHEN, 2018). As informações foram geradas a partir de parcerias com os profissionais que conduzem as pesquisas envolvendo os dados do ECR-PF, uma vez que as informações geradas necessitaram da análise e interpretação de um profissional de saúde. As minerações realizadas para este trabalho buscaram descrever padrões para ocorrência de fístulas, e dois artigos foram gerados visando responder as seguintes questões de pesquisa:

- É possível prever a ocorrência de fístulas pós palatoplastia primária em pacientes com fissura transforame unilateral?

- Qual tipo de algoritmo de aprendizado de máquina tem o melhor desempenho nesta tarefa de predição, considerando como medidas de desempenho as métricas de acurácia, precisão, cobertura e f-measure?
- A utilização de técnicas de mineração de dados pode evidenciar algum conhecimento útil sobre fatores associados à ocorrência de fístulas pós palatoplastia primária em pacientes com fissura transforame unilateral (FTU)?
- Que tipos de descobertas as diferentes classes de tarefas de mineração de dados (classificação e associação), puderam evidenciar?

Deste modo, foram propostas duas tarefas de pesquisa que são: a comparação de algoritmos de aprendizado de máquina para a predição de ocorrência de fístulas pós palatoplastia primária em pacientes com fissura transforame unilateral (FTU) e a utilização de técnicas de mineração de dados para a descoberta de conhecimento sobre fatores associados à ocorrência de fístulas pós palatoplastia primária nesses pacientes.

Para comparar diferentes algoritmos de aprendizado de máquina para a predição de ocorrência de fístulas pós palatoplastia primária em pacientes com fissura transforame unilateral foi proposta uma investigação com diferentes amostras de pacientes participantes do ECR-PF sendo que todos apresentaram FTU, sem síndromes, operados por quatro cirurgiões, usando duas técnicas na palatoplastia primária (Furlow e Langenbeck). Foi selecionado um subconjunto de dados de 371 pacientes que apresentaram dados de interesse deste estudo com informações sobre ocorrência de fístula, conforme o acervo anonimizado consultado. A linguagem de programação Python foi utilizada para implementação e comparação de 6 diferentes algoritmos classificadores (Árvore de decisão (decision Tree), K-Vizinhos Mais Próximos (k-Nearest Neighbor – k-NN), Regressão Logística (logistic Regression), Naive Bayes Gaussiano (Gaussian Naive Bayes), Máquinas de Vetores-suporte (Support Vector Machine) e Floresta Aleatória (Random Forest)). Os algoritmos foram comparados considerando as métricas de acurácia, precisão, cobertura e f-measure.

Para desenvolver um processo de descoberta de conhecimento sobre fatores associados à ocorrência de fístulas pós palatoplastia primária em pacientes com fissura transforame incisivo unilateral (FTIU) foi proposto um estudo com a utilização de diferentes técnicas de mineração de dados aplicadas a um subconjunto de dados de 222 pacientes com informações sobre ocorrência de fístulas (destacase que foram descartados para análise pacientes com dados incompletos para qualquer uma das variáveis estudadas). Ao analisar a base de dados junto com profissionais da saúde se observou que nem todos os protocolos do estudo foram totalmente preenchidos pelos profissionais da área médica. Foram realizadas duas tarefas de mineração de dados (classificação (J48) e associação (a priori)), utilizando o software WEKA.

Os dois artigos apresentados a seguir dão suporte à hipótese principal deste trabalho de que o uso de algoritmos de aprendizado de máquina (em tarefas de classificação e associação) em um contexto de mineração de dados são capazes de evidenciar correlações e padrões "escondidos" em bases de dados relacionadas às fissuras labiopalatinas, de modo a gerar informação útil (insights) referentes, principalmente, à ocorrência de fístulas após a palatoplastia primária:

Artigo 1: Aplicação de mineração de dados para descoberta de conhecimento sobre fatores associados à ocorrência de fístulas após palatoplastia;

Artigo 2: Predição da ocorrência de fístulas nas fissuras labiopalatinas utilizando técnicas de aprendizado de máquina.

5

Artigos

5 ARTIGOS

A seguir são apresentados os dois artigos (submetidos) que compõem esta tese.

5.1 ARTIGO 1

PREDIÇÃO DA OCORRÊNCIA DE FÍSTULAS NAS FISSURAS LABIOPALATINAS UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

O presente artigo foi escrito e submetido de acordo com as instruções e normas da Revista de Saúde Digital e Tecnologias Educacionais (ISSN: 2525-9563).

Autores:

Patrick Pedreira Silva

Programa de Pós-Graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Elvio Gilberto da Silva

Programa de Pós-Graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Vinicius Santos Andrade

Programa de Pós-Graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Telma Vidotto de Sousa Brosco

Departamento de Cirurgia Plástica. Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Maria Inês Pegoraro Krook

Departamento de Fonoaudiologia da Faculdade de Odontologia de Bauru FOB-USP, Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Jeniffer de Cássia Rillo Dutka

Departamento de Fonoaudiologia da Faculdade de Odontologia de Bauru FOB-USP,
Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São
Paulo (USP), Bauru-SP



Predição da ocorrência de fístulas nas fissuras labiopalatinas utilizando técnicas de aprendizado de máquina

Patrick Pedreira Silva, Elvio Gilberto da Silva, Vinicius Santos Andrade, Telma Vidotto de Sousa Brosco, Maria Inês Pegoraro Krook, Jeniffer de Cássia Rillo Dutka

Resumo

Introdução: O aumento da quantidade de dados armazenados em prontuários eletrônicos de pacientes amplia a possibilidade de obtenção de informações importantes no apoio ao processo decisório dos profissionais de saúde. Entretanto, muitas vezes, o grande volume de dados dificulta o gerenciamento e análise das informações obtidas, demandando processos automatizados para a manipulação de tais dados. **Objetivo:** O estudo propõe a comparação de algoritmos de aprendizado de máquina para a predição de ocorrência de fístulas pós palatoplastia primária em pacientes com fissura transforame unilateral (FTU). **Material e métodos:** De uma amostra de 466 pacientes participantes em um estudo clínico randomizado com FTU, sem síndromes, operados por quatro cirurgiões, usando duas técnicas na palatoplastyia primária (Furlow e Langenbeck), selecionou-se um subconjunto de dados de 372 pacientes com informações sobre ocorrência de fístula. A linguagem de programação Python foi utilizada para implementação e comparação de 6 diferentes algoritmos classificadores (Árvore de decisão (decision Tree), K-Vizinhos Mais Próximos (k-Nearest Neighbor – k-NN), Regressão Logística (logistic Regresion), Naive Bayes Gaussiano (Gaussian Naive Bayes), Máquinas de Vetores-suporte (Suport Vector Machine) e Floresta Aleatória (Random Forest)). **Resultados:** Os algoritmos foram comparados considerando as métricas de acurácia, precisão, cobertura e f-measure. Com relação à acurácia o melhor desempenho preditivo foi obtido pelo algoritmo Naive Bayes Gaussiano. Em termos de precisão o melhor desempenho foi do algoritmo de Árvore de Decisão. Considerando-se a cobertura e a f-measure o melhor foi o algoritmo de Máquina de Vetor Suporte. Entretanto, quando se considera a importância de predizer os casos de insucesso, já que isso pode influenciar no tratamento dos pacientes, os algoritmos que se destacam são Máquina de Vetor Suporte com uma precisão de 0,97 e a Árvore de Decisão com valor de cobertura de 0,67 e f-measure de 0,69. **Conclusão:** Especificamente com relação à base de dados analisada quando se considera a capacidade de predição (correta e do maior número de casos) o algoritmo de melhor desempenho foi o de Máquina de Vetor Suporte, cuja métrica de f-measure dentro da classe de insucessos foi a mais alta.

1. Introdução

Um dos principais objetivos da cirurgia primária do palato para a correção da fissura labiopalatina é a reconstrução bem-sucedida da cinta muscular dos músculos elevadores do

palato, de forma a propiciar um mecanismo velofaríngeo funcional para o desenvolvimento de fala e para o funcionamento da orelha média, além de evitar o comprometimento do crescimento facial¹. A ocorrência de fístulas oronasais residuais é uma das complicações possíveis após a cirurgia primária do palato^{2,3,4,5,6}. A definição de fístula, conforme reportado por Brosco⁶, é uma falha na cicatrização ou ruptura da cirurgia primária do palato⁷ e sua incidência pode variar entre 0%⁸ e 78%⁹. Para prevenir estas complicações cirúrgicas é importante entender melhor os fatores associados à ocorrência de fístulas e esta não é uma tarefa fácil, quando se trata de instituições que lidam com uma grande casuística.

A chamada era da informação caracteriza-se pela crescente expansão do volume de dados gerado e armazenado, fenômeno que também se reflete na área de saúde, ampliando a possibilidade de obtenção de informações importantes para o processo decisório do tratamento¹⁰. Dados disponibilizados em prontuários de pacientes são gerados com rapidez, acumulando um grande volume de informações de difícil utilização ao considerar-se uma análise manual. Processos automatizados e mais sofisticados para a manipulação de tais dados, por sua vez, ampliam a possibilidade da identificação e gerenciamento de informações, muitas vezes ocultas nos prontuários.

É exatamente neste contexto, de superabundância de dados, que surgiu a mineração de dados como um processo sistemático, interativo e iterativo, de preparação e extração de conhecimentos a partir de grandes bases de dados¹⁰. Neste cenário, este trabalho compara técnicas de aprendizado de máquina para identificar o melhor algoritmo para prever a ocorrência de fístulas após a palatoplastia em pacientes com fissura labiopalatina.

2. Métodos

O presente estudo foi aprovado pelo Comitê de Ética em Pesquisa do INFORMAÇÃO REMOVIDA, (parecer 1.753.467). A amostra refere-se a um conjunto de dados de prontuários de 371 pacientes com fissura labiopalatina unilateral não associada a outras malformações, que apresentaram fístula de palato após terem sido submetidos à cirurgia primária do palato.

Variáveis estudadas

No gerenciamento da fissura labiopalatina um bom resultado (sucesso do tratamento) ocorre na ausência de fístula e de disfunção velofaríngea com um mínimo de comprometimento do crescimento do terço médio da face. A ocorrência de fístula no palato, em região posterior ao forame incisivo, foi a complicação clínica pós-palatoplastia primária estudada no presente estudo.

Para classificação de fístula dos dados extraídos dos prontuários dos pacientes o protocolo descrito por Brosco⁶ foi usado, dividindo os pacientes em quatro grupos: grupo Sem Fístula (SEMFI); grupo Com Fístula na Região Pré-Forame Incisivo (PREFI); grupo Com Fístula na Região Pós-Forame Incisivo (PROFI); e grupo Com Fístula Envolvendo tanto a Região Pré quanto Pós-Forame Incisivo (PREPO). Para o presente estudo, a ocorrência de fístula em regiões posteriores ao forame incisivo (POSFI e PREPO), particularmente, foi interpretada como indicativa de complicação do tratamento cirúrgico primário. A fístula na região anterior ao forame incisivo (PREFI) não foi considerada uma complicação.

A pergunta norteadora da investigação envolveu definir quais os melhores algoritmos para prever a ocorrência ou não das fístulas POSFI e PREPO após a cirurgia primária do palato em indivíduos com fissura labiopalatina unilateral. Para verificar, portanto, se algumas variáveis podiam ou não ser usadas como preditoras da ocorrência de fístula, coletou-se nos prontuários estudados (Tabela 1) os seguintes dados: a) a técnica utilizada na cirurgia primária para correção da fissura do lábio (queiloplastia primária); b) a técnica utilizada na cirurgia primária para correção da fissura de palato (palatoplastia primária); c) o tempo de duração da palatoplastia em minutos; d) a idade do paciente quando submetido à palatoplastia; e) uso de incisão relaxante; f) uso de retalho de vômer; g) ocorrência e localização da fístula no palato (SEMFI, PREFI, POSFI ou PREPO); h) a ocorrência de infecção após a palatoplastia no local da cirurgia ou em outro local; i) a ocorrência de vômito ou tosse no pós-operatório da palatoplastia.

A casuística do presente estudo fez parte de um estudo prospectivo randomizado, cujo objetivo foi comparar os resultados de fala e do crescimento facial de indivíduos com fissura labiopalatina unilateral operados de lábio e palato nas seguintes condições: técnica de Millard ou Spina para a cirurgia de lábio e Furlow ou von Langenbeck para a de palato. Durante a palatoplastia o cirurgião poderia ou não ter usado incisão relaxante (para diminuir a tensão na área da sutura) e usado retalho de vômer (para liberar tecido para corrigir fendas mais amplas). As fístulas POSFI e PREPO foram combinadas e tratadas como casos com complicação após a cirurgia. As fístulas PREFI foram combinadas aos casos SEMFI e tratadas como casos sem complicação após a cirurgia.

Procedimentos aplicados

Foram utilizados seis diferentes algoritmos classificadores incluindo: Árvore de Decisão (Decision Tree), K-Vizinhos Mais Próximos (K-Nearest Neighbor – K-NN), Regressão Logística (Logistic Regression), Naive Bayes Gaussiano (Gaussian Naive Bayes), Máquinas de Vetores-Suporte (Support Vector Machine) e Floresta Aleatória (Random Forest). O algoritmo

de Árvore de Decisão (AD) gera uma árvore de regras que permite fazer predições por meio da relação entre as variáveis. O K-Vizinhos Mais Próximos (KNN) realiza a classificação de cada variável com base no conhecimento de todas as classes de variáveis consideradas e, para tal, utiliza métricas específicas (como por exemplo, a distância euclidiana) para comparar as distâncias entre as diferentes variáveis (objetos). A Regressão Logística (RL) procura encontrar uma função ou conjunto de funções que discrimine grupos definidos de variáveis pelo atributo classe, visando minimizar erros de classificação. O Naive Bayes Gaussiano (NBG) se baseia na aplicação do Teorema de Bayes com a hipótese ingênua (naive) de independência entre cada par de fatores (variáveis), fornecendo predições associadas a probabilidades da ocorrência em estudo. As Máquinas de Vetores-Suporte (SVM) geram um hiperplano a partir dos pontos de classe, o hiperplano chamado Vetores de Suporte é utilizado como margem de separação entre classes. Neste estudo as variáveis de interesse foram tratadas como os atributos e os algoritmos foram implementados utilizando-se a biblioteca Scikit Learn do Python. Python é uma linguagem de programação que pode ser utilizada para o aprendizado de máquina, análise e tarefas de mineração de dados. Esta linguagem permite visualizar e explorar os conjuntos de dados (variáveis de interesse), aplicar filtros e executar tarefas de mineração tais como classificação, agrupamentos e associações.

Considerando-se uma tarefa de mineração típica, o estudo foi dividido em quatro etapas: pré-processamento de dados, extração de características, classificação e descrição de resultados. O procedimento foi realizado considerando-se como resultado primário a ocorrência de fistula após a palatoplastia. O pré-processamento foi realizado de maneira manual organizando-se os dados dos prontuários em arquivo “.XLS” (planilha de Excel® Microsoft Office). Os dados levantados pelos pesquisadores foram convertidos em dataframes, que são estruturas similares às tabelas. Um programa em Python foi implementado de forma a permitir a análise dos dados e a comparação entre os algoritmos, tendo-se como bases métricas a acurácia, a precisão, a cobertura e a f-measure.

Todas essas métricas são calculadas com valores no intervalo entre 0 e 1 e estão baseadas no conceito de matriz de confusão que se trata de uma matriz com dimensões proporcionais à quantidade de classes que se deseja prever. A matriz de confusão é voltada para modelos de classificação e tem como objetivo indicar a quantidade de Falsos Positivos (FP) e Falsos Negativos (FN); e de Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN), que vão dar suporte aos cálculos das métricas indicadas.

Especificamente neste estudo, os modelos foram utilizados para classificar a ocorrência ou não das fístulas (consideradas ou não complicações pós-cirúrgicas), tendo

como base os dados de uma matriz de dimensão 2X2 (Quadro 1) por considerar apenas duas classes: SEM (pacientes que não tiveram fístula consideradas complicações pós-cirúrgicas) e COM (aqueles que tiveram fístula consideradas complicações, pois ocorreram após o forame incisivo). Após treinar o modelo, os dados da matriz de confusão foram utilizados como base para o cômputo das métricas. Para entendimento da matriz e das métricas foram consideradas as definições seguintes.

Quadro 1. Matriz de confusão que relaciona as classes consideradas neste estudo

		Classe esperada	
		COM	SEM
Classe prevista	OM C	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	EM S	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: Pedreira-Silva P, Silva EG da, Andrade VS, Brosco TVS, Prearo GA, Krook MIP, Dutka JCR.

Quando o modelo prevê um caso positivo corretamente, significa que ele previu a ocorrência de fístula consideradas complicações (COM) corretamente, portanto aqui tem-se um caso de Verdadeiro Positivo (VP). Ao contrário, quando o modelo prevê um caso positivo incorretamente, significa que ele previu a ocorrência de fístula (COM) incorretamente, portanto, aqui, quando na verdade não ocorreu (SEM), tem-se um caso de Falso Positivo (FP).

O oposto também pode ocorrer. Quando o modelo indica que não houve fístula consideradas complicações pós-cirúrgicas (SEM), mas na verdade ocorreu (COM), trata-se de um caso de Falso Negativo (FN), e quando ele indica que não ocorreu (SEM) e a previsão está correta, ocorre o Verdadeiro Negativo (VN). Uma vez identificados os valores de VP, VN, FP e FN é possível calcular as métricas de precisão, cobertura, acurácia e *f-measure*.

A métrica de precisão refere-se à capacidade do modelo de evitar falsos positivos no processo de classificação. Em outras palavras a precisão indica o percentual de acertos dentro de uma classe específica (dentre os elementos preditos como sendo de uma determinada classe (COM / SEM), quantos realmente são desta classe?)

$$precisão = \frac{VP}{VP + FP}$$

A métrica de cobertura é a proporção entre as instâncias classificadas corretamente (VP) e o total de instâncias da amostra daquela classe (VP + FN) (quantas instâncias de uma determinada classe (COM / SEM) foram classificadas corretamente?).

$$cobertura = \frac{VP}{VP + FN}$$

A acurácia corresponde à proporção entre as instâncias que foram corretamente previstas, sejam elas verdadeiro positivo (VP) ou verdadeiro negativo (VN). Em outras palavras ela indica o percentual de acertos, considerando a totalidade dos dados usada na fase de teste do algoritmo (quantas instâncias, na totalidade da amostra, foram classificadas corretamente considerando as duas classes (COM / SEM) simultaneamente?).

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

A *f-measure* é uma medida que combina precisão e cobertura, representando a média harmônica entre ambas. Valores próximos de 1 significam que a acurácia obtida é relevante, ou seja, os valores de VP, VN, FP, FN aferidos não apresentam grandes distorções. Também pode-se interpretar como uma medida de confiabilidade da acurácia.

$$f - measure = 2 * \frac{precisão * cobertura}{precisão + cobertura}$$

Em termos simples, altos valores de precisão e acurácia significam que o algoritmo retornou mais resultados relevantes que irrelevantes, enquanto alta cobertura significa que o algoritmo retornou a maioria dos resultados relevantes. Para o cálculo das métricas de classificação dos algoritmos, os dados foram particionados (em conjuntos de treinamento e teste). Normalmente, a amostra é separada nessas duas categorias, assim, o conjunto de treinamento é utilizado para a criação dos modelos de classificação e o conjunto de teste é utilizado para verificar sua performance, conforme as métricas. O benefício de separar a amostra em subconjuntos de treinamento e de teste é que o modelo pode ser testado para dados nunca antes vistos durante a fase de treinamento, e, assim, garantir a sua capacidade de generalização, evitando que o modelo seja excessivamente sensível aos dados com que foi treinado.

Neste estudo o particionamento dos dados foi realizado usando a validação cruzada (*K-fold cross-validation*), em que o conjunto das instâncias da base de dados é dividido aleatoriamente em k partições de mesmo tamanho (subconjuntos das amostras). O intuito foi criar diferentes conjuntos de treino e teste, treinando o modelo e medindo a sua performance frente a dados diversos. Nesse caso, ao invés de usar apenas um conjunto de teste para validar o modelo, foram utilizados N outros a partir dos mesmos dados, ou seja, a cada iteração do algoritmo, uma partição é utilizada como teste e as demais para o

treinamento. Esse processo é repetido tantas vezes quanto for o número de partições. Neste estudo foi utilizado o método de validação cruzada com amostra sendo dividida em 5 ($k=5$) partes (aleatórias, com cerca de 74 instâncias cada). Para cada uma dessas partições, o modelo usa quatro partes ($K-1$) para treinar, enquanto usa uma parte para testar. Ao final do processo, quando o modelo iterar/treinar cinco vezes, as métricas de performance são estimadas, com base na média de todos os treinos realizados.

Para o algoritmo de árvores de decisão (AD), como critério para fazer as partições dos nós foi utilizado o índice Gini¹¹. Além disso, foi definido um parâmetro que visou indicar o número mínimo de instâncias por folhas (20). Esses mesmos critérios foram definidos para o método da Floresta Aleatória (RF), adicionalmente, considerou-se 10 estimadores como parâmetro. Os demais algoritmos (RL, KNN, NBG, SVM) foram utilizados na sua versão padrão (default) do Scikit Learn.

3. Resultados

As informações sobre a ocorrência de algum tipo de fístula foram identificadas nos prontuários dos 371 pacientes que constituíram a amostra do presente estudo. Os dados levantados nos prontuários e usados no treinamento dos algoritmos foram agrupados conforme as variáveis listadas na Tabela 1, com seus respectivos valores. No grupo de 371 prontuários, 150 (40,4%) pertenciam aos pacientes do sexo feminino e 221 (59,6%) do sexo masculino. Na amostra estudada, 312 (84,1%) pacientes foram sub-agrupados como aqueles que não tiveram fístula consideradas complicações pós-cirúrgicas (SEMFI ou PREFI), enquanto 59 (15,9%) foram sub-agrupados como aqueles que tiveram fístula consideradas complicações, pois ocorreram após o forame incisivo (POSFI ou PREPO). A distribuição dos dados em porcentagem (%) e quantidades, de acordo com a ocorrência de fístula, foi a seguinte: SEMFI (grupo sem fístula) (N=245; 66%); PREFI (grupo com fístula na região pré-forame incisivo) (N=67; 18,1%); POSFI (grupo com fístula na região pós-forame incisivo) (N=32; 8,6%); e PREPO (grupo com fístula envolvendo tanto a região pré- quanto pós-forame incisivo) (N=27; 7,3%).

Quanto à técnica utilizada na queiloplastia primária, verificou-se que, do total de 371 (100%) pacientes, 193 (52,0%) receberam a técnica de Millard, enquanto 178 (48,0%) a de Spina. Quanto à técnica utilizada na palatoplastia primária, do total de 371 (100%) pacientes, 177 (47,7%) receberam a técnica de Furlow, enquanto 194 (52,3%) a de von Langenbeck. Quanto à ocorrência de fístula, dos 177 (100%) que receberam a técnica de Furlow, 115 (64,9%) foram classificados como SEMFI, 26 (14,7%) PREFI, 20 (11,3%) POSFI e 16 (9,1%) PREPO e dentre os 194 (100%) pacientes que receberam a técnica de Langenbeck,

130 (67%) foram classificados como SEMFI, 41 (21,1%) como PREFI, 12 (6%) como POSFI e 11 (5,9%) como PREPO.

A idade média dos pacientes na palatoplastia primária foi 12,9 meses ($\sigma=3,2$; $\text{min}=9,0$; $\text{max}=22,0$). Quando os dados foram agrupados de acordo com a presença de fístulas consideradas como complicação cirúrgica (POSFI+PREPO=59), a idade média foi 12,2 meses ($\sigma: 3,26$; $\text{Min}: 9,0$; $\text{Max } 20,0$) comparada à média de 13,1 meses ($\sigma: 3,26$; $\text{Min } 9,0$; $\text{Max } 22,0$) para o grupo considerado sem complicação cirúrgica (SEMF+PREFI=312). A Tabela 1 sumariza as variáveis estudadas nas diversas categorias.

Tabela 1. Definição das variáveis de interesse e seus respectivos valores para este estudo

Variáveis	Valores
Técnica cirúrgica na queiloplastia	Millard: N=193, Spina: N=178
Técnica cirúrgica na palatoplastia	Furlow: N=177, von Langenbeck: N=194
Idade na palatoplastia (Meses)	Média:12,9; σ :3,2; Min:9,0; Max:22,0
Incisão relaxante	Sem: N=91, Unilateral: N=80, Bilateral: N=200
Retalho de Vômer	Sim: N=174, Não: N=197
Duração da palatoplastia (Minutos)	Média: 61,8; σ :23,3; Min:25; Max:140
Infecção na palatoplastia	Não: N=349, No Local da Cirurgia: N=9, em Outro Local: N=13
Vômito após a palatoplastia	Sim: N=285, Não: N=86
Tosse após a palatoplastia	Sim: N=349, Não: N=22
Febre	Sim: N=261, Não: N=110
Ocorrência de fístula*	SEMF: N=245, PREFI: N=67, POSFI: N=32, PREPO: N=27

Fonte: Pedreira-Silva P, Silva EG da, Andrade VS, Brosco TVS, Prearo GA, Krook MIP, Dutka JCR.

Com relação ao tempo de duração da palatoplastia em minutos, a média para os 371 casos foi de 61,8 minutos ($\sigma: 23,3$; $\text{Min}: 25,0$; $\text{Max}: 140,0$). Especificamente para cada grupo, os 312 casos SEMFI+PREFI ("SEM") apresentaram duração média da palatoplastia de 59,44 minutos ($\sigma: 22,22$; $\text{Min}: 25,0$; $\text{Max}: 125,0$), enquanto os 59 casos POSFI+PREPO ("COM") apresentaram duração média da palatoplastia de 74,57 minutos ($\sigma: 23,37$; $\text{Min } 25,0$; $\text{Max } 140,0$). Há indícios, portanto, que cirurgias mais demoradas estão associadas a piores

resultados. Observando a classe “SEM” a maior parte das cirurgias (67,62%) teve duração de no máximo 60 minutos. Já na classe “COM” a maior parte das cirurgias (58,33%) durou mais de 60 minutos.

Foram aplicadas as técnicas de aprendizado de máquina (TAM) considerando-se todas as variáveis (Tabela 1), estabelecendo-se médias das medidas de acurácia, de precisão, de cobertura e de f-measure para cada algoritmo de interesse (RL, KNN, AD, NBG, SVM, RF), utilizando-se o 5-fold cross-validation (k=5), conforme Tabela 2.

Tabela 2. Resultados das métricas para os algoritmos (valores no intervalo de 0 a 1), considerando-se as médias obtidas para ambas as classes (SEMFI+PREFI e POSFI+PREPO) nas diferentes técnicas de aprendizado de máquina

Média	RL	KNN	AD	NBG	SVM	RF
Acurácia	0,88*	0,85	0,82	0,88*	0,84	0,84
Precisão	0,85	0,78	0,78	0,78	0,95*	0,42
Cobertura	0,70	0,65	0,80*	0,74	0,77	0,50
f-measure	0,75	0,68	0,79	0,76	0,83*	0,46

* valor mais alto

Fonte: Pedreira-Silva P, Silva EG da, Andrade VS, Brosco TVS, Prearo GA, Krook MIP, Dutka JCR.

Observa-se que a acurácia média para cada tipo de técnica de aprendizado de máquina variou entre o mínimo de 0,82 e o máximo de 0,88. A precisão variou entre 0,42 e 0,95. A cobertura variou entre 0,50 e 0,80 e a f-measure variou entre 0,46 e 0,83. As técnicas de NBG e RL foram as que apresentaram acurácia mais alta (0,88) ao identificar corretamente a maior parte dos casos sem e com fístulas consideradas complicações pós-operatórias, enquanto a técnica SVM teve maior precisão média (0,95) e maior f-measure média (0,83). A melhor cobertura média foi da técnica AD (0,80).

Buscando identificar os valores que indicam as melhores TAM para prever a ocorrência de fístulas consideradas complicações pós-palatoplastia (POSFI e PREPO) os dados foram analisados de acordo com a ocorrência das complicações. Os valores de precisão, de cobertura e de f-measure foram estabelecidos para cada uma das 6 TAM (RL, KNN, AD, NBG, SVM, RF), e são apresentados nas Tabela 3.

Tabela 3. Resultados das métricas de precisão (pre), de cobertura (cob) e de f-measure (f-m) para os algoritmos de cada técnica de aprendizagem de máquina para o grupo com complicações (POSFI e PREPO)

	RL	KNN	AD	NBG	SVM	RF
pre	0,81	0,68	0,62	0,64	0,97*	0,00
cob	0,42	0,32	0,67*	0,53	0,53	0,00
f-m	0,55	0,43	0,65	0,58	0,69*	0,00

* valor mais alto

Fonte: Pedreira-Silva P, Silva EG da, Andrade VS, Brosco TVS, Prearo GA, Krook MIP, Dutka JCR.

Na Tabela 4 são apresentados os valores das métricas para o grupo considerado sem complicações (SEMFI e PREFI). Observa-se que a precisão para identificar o grupo com complicações (POSFI e PREPO) é na maioria das vezes menor do que a precisão para identificar o grupo sem complicações (SEMFI e PREFI). Nota-se, particularmente, que a técnica SVM teve a maior precisão deste estudo (0,97) e também melhor f-measure (0,69) para prever os 59 casos com complicações. A técnica que apresentou a melhor cobertura destes 59 casos foi a AD (0,67).

Tabela 4. Resultados das métricas de precisão (pre), de cobertura (cob) e de f-measure (f-m) para os algoritmos de cada técnica de aprendizagem de máquina para o grupo sem complicações (SEMFI e PREFI)

	RL	KNN	AD	NBG	SVM	RF
pre	0,90	0,88	0,94*	0,91	0,92	0,84
cob	0,98	0,97	0,92	0,94	1,00*	1,00
f-m	0,94	0,92	0,93	0,93	0,96*	0,91

* valor mais alto

Fonte: Pedreira-Silva P, Silva EG da, Andrade VS, Brosco TVS, Prearo GA, Krook MIP, Dutka JCR.

Conforme a Tabela 4, com referência ao grupo sem complicações (SEMFI e PREFI) o algoritmo com melhor precisão é a AD (0,94), já o SVM apresenta os melhores valores para cobertura (1,00) e f-measure (0,96). De modo geral, os algoritmos apresentam um valor médio de 0,93 para a métrica de f-measure.

Discussão

A oportunidade de adotar a mineração dos dados de pacientes com fissura labiopalatina pode oportunizar um melhor entendimento das especificidades em relação à ocorrência ou não de fístulas consideradas complicações após a palatoplastia primária. Um melhor entendimento dos aspectos clínicos e técnicos relacionados ao gerenciamento da fissura pode favorecer tanto a prevenção de complicações quanto o processo de definição

da melhor conduta de tratamento. Neste estudo específico, a escolha das variáveis de interesse (Tabela 2) juntamente com diferentes algoritmos de mineração (RL, KNN, AD, NBG, SVM, RF) ofereceu oportunidades aos profissionais de saúde e de informática de identificarem as técnicas computacionais que melhor permitem prever a ocorrência de fístulas pós-palatoplastia primária em pacientes com fissura labiopalatina.

Da mesma forma que em outros estudos^{12,13,14,15,16,17} esta investigação optou pela comparação de diferentes algoritmos de mineração de dados, buscando verificar a técnica de aprendizagem de máquina que melhor identificasse as diferentes classes associadas à amostra (base de dados) considerada. Conforme observado na Tabela 2 o melhor desempenho preditivo, quando se considera a métrica de acurácia, foi obtido pelos algoritmos NBG e RL. Em termos de precisão média (considerando ambos os casos com e sem fístulas) o melhor desempenho foi do algoritmo SVM, que também apresentou a melhor f-measure. Ao se considerar a cobertura, o melhor desempenho médio foi do algoritmo de AD. Em termos de desempenhos médios em cada uma das métricas, as técnicas de KNN e RF, foram as de piores resultados para a base de dados do estudo.

Quando se considera a importância de prever os casos com complicações (59 do total de 371), já que isso pode influenciar o tratamento dos pacientes, os algoritmos que se destacam, conforme a Tabela 3, são (SVM) com uma precisão de 0,97 e a AD com valor de cobertura de 0,67. O algoritmo SVM também é o melhor em termos de f-measure (0,69), mostrando o seu melhor balanceamento em termos de precisão e cobertura. Os dados da Tabela 4 mostram que os algoritmos, de um modo geral, apresentam um bom desempenho quando se trata da predição de casos sem fístulas pois, em média, eles apresentam um valor 0,93 para a métrica de f-measure.

Reconhece-se, ainda, que este estudo oferece somente uma perspectiva pontual da realidade, por meio das análises preliminares de técnicas de mineração de dados na base considerada, indicando algoritmos que podem ser mais explorados em estudos futuros. Quando se considera apenas prever a ocorrência de fístulas consideradas complicações, o algoritmo SVM apresenta-se como um bom candidato. Pelos resultados apontados por este estudo e por suas características inerentes (evidenciando possíveis regras de classificação) a técnica de AD poderá servir como base de novas investigações sobre como e quais fatores estão associados aos resultados da palatoplastia. Do ponto de vista dos algoritmos de aprendizados de máquina, existe sempre a necessidade de validação dos achados por profissionais da área de saúde envolvidos no tratamento da fissura labiopalatina, caracterizando-se uma ação interdisciplinar com parceria da saúde e ciência da computação.

É importante considerar neste estudo as possíveis limitações como, por exemplo, o fato de a amostra estar desbalanceada (59 casos com complicações comparados à 312 sem complicações). No entanto, as amostras refletem a situação real em casos com complicações, uma vez que fístulas ocorrem em menor quantidade do que casos sem complicações, mantendo-se neste estudo a proporção natural dos dados. Essa presença de classes majoritárias com frequência muito maior que as outras classes minoritárias, faz com que os algoritmos tenham uma tendência para responder bem para as classes majoritárias em detrimento das minoritárias (como observado com algumas técnicas nas Tabelas 2, 3 e 4). Em trabalhos futuros, o experimento poderá ser repetido usando técnicas de resampling aleatório (undersampling ou oversampling), isto é, uma re-amostragem dos exemplos de treinamento de forma a gerar conjuntos balanceados¹⁸.

A utilização da validação cruzada (cross-validation) para treinamento e teste, procurou evitar algum tipo de viés de ajuste aos dados. Qualquer método de aprendizado de máquina pode sofrer overfitting que é observado quando o modelo produz um classificador que se ajusta bem aos dados de treinamento, mas não consegue generalizar o conhecimento aprendido, não obtendo um bom desempenho nos dados de teste. É como se o classificador decorasse os dados do treinamento. No estudo foram observados valores de acurácia acima de 0,82 em todos os algoritmos e também valores médios relativamente altos de precisão e cobertura, indicando que os algoritmos desempenharam bem a tarefa de classificação, sobretudo, nos casos sem complicações (SEMI e PREFI), apresentando maior dificuldade para classificar pacientes com fístulas do tipo PREPO ou POSFI.

O intuito deste experimento não foi, necessariamente, automatizar o processo de classificação e nem de definir como os diferentes fatores estão associados à presença de fístulas, mas sim, comparar algoritmos e evidenciar o potencial da utilização da abordagem da aprendizagem de máquina na tarefa de predição para o problema escolhido. Um futuro estudo, portanto, pode ajudar a encontrar melhores modelos que permitam, por exemplo, o desenvolvimento de softwares de apoio à decisão clínica.

4. Conclusão

Este trabalho comparou o desempenho de diferentes algoritmos de aprendizado de máquina em problemas de classificação de fístulas após palatoplastia. Para isso, diferentes métricas propostas na literatura para avaliação de classificadores foram utilizadas. Para a realização dos experimentos, foram utilizados os algoritmos Árvore de decisão (AD), K-Vizinhos Mais Próximos (KNN), Regressão Logística (RL), Naive Bayes Gaussiano (NBG),

Máquinas de Vetores-suporte (SVM) e Floresta Aleatória (RF). A análise dos dados revelou que fatores associados às complicações pós-operatórias (infecção, vômito, tosse e febre) bem como características associadas à cirurgia em si (duração, técnicas cirúrgicas, retalho de Vômer, incisão relaxante) e ao paciente (idade na época da palatoplastia), podem ajudar a prever o sucesso ou insucesso da palatoplastia com relação à ocorrência de fístulas consideradas complicações. Quando se considera a capacidade de predição (correta e do maior número de casos) o algoritmo de melhor desempenho foi o de Máquina de Vetores-Suporte (SVM), cuja métrica de f-measure dentro da classe de insucessos foi a mais alta.

5. Referências

1. Dutka JDCR; Krook MIP. Avaliação e tratamento das disfunções velofaríngeas. In: Marchesan IQ, Justino H, et al, editores. Tratado de especialidades em Fonoaudiologia. São Paulo: Grupo Gen-Editora Roca Ltda., 2014. cap. 40, p.363-368. ISBN 9788527726412. ISBN 9788527726412.
2. Hardwicke JT, Landini G, Richard BM. Fistula incidence after primary cleft palate repair: a systematic review of the literature. *Plast Reconstr Surg.* 2014 Oct;134(4):618e-27e. doi: 10.1097/PRS.0000000000000548. PMID: 25357056.
3. de Agostino Biella Passos V, de Carvalho Carrara CF, da Silva Dalben G, Costa B, Gomide MR. Prevalence, cause, and location of palatal fistula in operated complete unilateral cleft lip and palate: retrospective study. *Cleft Palate Craniofac J.* 2014 Mar;51(2):158-64. doi: 10.1597/11-190. Epub 2013 Apr 16. PMID: 23586365.
4. Deshpande GS, Campbell A, Jagtap R, Restrepo C, Dobie H, Keenan HT, Sarma H. Early complications after cleft palate repair: a multivariate statistical analysis of 709 patients. *J Craniofac Surg.* 2014 Sep;25(5):1614-8. doi: 10.1097/SCS.0000000000001113. PMID: 25148623.
5. Aslam M, Ishaq I, Malik S, Fayyaz GQ. Frequency of oronasal fistulae in complete cleft palate repair. *J Coll Physicians Surg Pak.* 2015 Jan;25(1):46-9. PMID: 25604369.
6. Brosco TVDS. Fístula de palato após reparo da fissura labiopalatina em um estudo clínico randomizado. Bauru. Tese de Doutorado - USP; 2017.
7. Muzaffar AR, Byrd HS, Rohrich RJ, Johns DF, LeBlanc D, Beran SJ, Anderson C, Papaioannou aA A. Incidence of cleft palate fistula: an institutional experience with two-stage palatal repair. *Plast Reconstr Surg.* 2001 Nov;108(6):1515-8. doi: 10.1097/00006534-200111000-00011. PMID: 11711920.
8. Dong Y, Dong F, Zhang X, Hao F, Shi P, Ren G, Yong P, Guo Y. An effect comparison between Furlow double opposing Z-plasty and two-flap palatoplasty on velopharyngeal closure. *Int J Oral Maxillofac Surg.* 2012 May;41(5):604-11. doi: 10.1016/j.ijom.2012.01.010. Epub 2012 Feb 15. PMID: 22340991.
9. Mak SY, Wong WH, Or CK, Poon AM. Incidence and cluster occurrence of palatal fistula after furrow palatoplasty by a single surgeon. *Ann Plast Surg.* 2006 Jul;57(1):55-9. doi: 10.1097/01.sap.0000205176.90736.e4. PMID: 16799309.
10. Ferrari DG, Castro LND. Introdução a mineração de dados. São Paulo: Saraiva; 2017. ISBN 8547200991.
11. Onoda M. Estudo sobre um algoritmo de árvores de decisão acoplado a um sistema de banco de dados relacional. Rio de Janeiro . Dissertação de Mestrado - Universidade Federal do Rio de Janeiro; 2001.

-
12. West D, Mangiameli P, Rampal R, West V. Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*. 2005;162(2):532-51.
 13. Teresinha M, Steiner A, Soma N, Shimizu T, Nievola J, José P, et al. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. *Gestão & Produção*. 2006;13.
 14. Junior ACB. Utilização de técnicas de data mining na detecção de outliers em auxílio à auditoria operacional com um estudo de caso com dados do sistema de informações hospitalares. Rio de Janeiro. Tese [Doutorado] – Universidade Federal do Rio de Janeiro; 2009.
 15. Kuretzki CH. Técnicas de mineração de dados aplicadas em bases de dados da saúde a partir de protocolos eletrônicos. Curitiba. Dissertação de Mestrado - Universidade Federal do Paraná; 2009.
 16. Martinez G, Bermúdez Y. Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial. *Revista Tecnura*. 2012;16.
 17. Carvalho D, Moser A, Silva V, Dallagassa M. Mineração de Dados aplicada à fisioterapia. *Fisioterapia em Movimento*. 2012; 25:595-605.
 18. Goldschmidt R, Passos E, Bezerra E. *Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações*. São Paulo: Elsevier Brasil, 2015. ISBN 8535278230.

5.2 ARTIGO 2

APLICAÇÃO DE MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO SOBRE FATORES ASSOCIADOS À OCORRÊNCIA DE FÍSTULAS APÓS PALATOPLASTIA

O presente artigo foi submetido e escrito de acordo com as instruções e normas da Revista Brasileira de Cirurgia Plástica (ISSN: 2177-1235)

Autores:

Patrick Pedreira Silva

Programa de Pós-Graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Elvio Gilberto da Silva

Programa de Pós-Graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Vinicius Santos Andrade

Programa de Pós-Graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Telma Vidotto de Sousa Brosco

Departamento de Cirurgia Plástica. Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Gabriela Aparecida Prearo

Programa de Pós-Graduação em Ciências da Reabilitação do Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São Paulo (USP), Bauru-SP

Maria Inês Pegoraro Krook

Departamento de Fonoaudiologia da Faculdade de Odontologia de Bauru FOB-USP,
Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São
Paulo (USP), Bauru-SP

Jeniffer de Cássia Rillo Dutka

Departamento de Fonoaudiologia da Faculdade de Odontologia de Bauru FOB-USP,
Hospital de Reabilitação de Anomalias Craniofaciais (HRAC), Universidade de São
Paulo (USP), Bauru-SP

APLICAÇÃO DE MINERAÇÃO DE DADOS PARA DESCOBERTA DE CONHECIMENTO SOBRE FATORES ASSOCIADOS À OCORRÊNCIA DE FÍSTULAS APÓS PALATOPLASTIA

Patrick Pedreira Silva, Elvio Gilberto da Silva, Vinicius Santos Andrade, Telma Vidotto de Sousa Brosco, Gabriela Aparecida Prearo, Maria Inês Pegoraro Krook, Jeniffer de Cássia Rillo Dutka

RESUMO

Introdução: Dados dos prontuários eletrônicos ampliam a obtenção de informações no processo decisório dos profissionais de saúde. Entretanto, o volume dificulta o gerenciamento e análise, demandando processos automatizados para a manipulação dos dados. Este estudo propõe a utilização de técnicas de mineração de dados para a descoberta de conhecimento sobre fatores associados à ocorrência de fístulas pós palatoplastia primária em pacientes com fissura transforame incisivo unilateral. Métodos: Foi utilizada uma base de 222 pacientes de um estudo clínico randomizado, sem síndromes, operados por quatro cirurgiões, usando duas técnicas de Furlow e Langenbeck, com informações sobre ocorrência de fístulas. Foram realizadas duas tarefas de mineração de dados (classificação (J48) e associação (a priori)), utilizando o software WEKA. Resultados: Cinco regras de uma árvore de decisão apontaram alguns dos fatores indicativos de fístulas nas cirurgias (infecção, tosse, hipernasalidade, cirurgião). A análise do modelo indica que ele classifica entre ausência e presença de fístulas corretamente 95,9%. As regras de associação geradas indicam que fatores combinados (ausência de infecção e de febre, ausência de hipernasalidade e ausência de resultado sugestivo de disfunção velofaríngea) estão relacionados à ausência de fístulas pós palatoplastia primária. Com relação

aos procedimentos cirúrgicos houve indícios de que a utilização da técnica de Furlow e de retalho de Vômer é mais frequente em pacientes com fístulas. Conclusão: As regras encontradas com alto grau de precisão e cobertura indicam que sintomas pós cirurgia (infecção e tosse), teste de hipernasalidade, cirurgião e a técnica cirúrgica podem ser preditores de fístulas após palatoplastia primária.

INTRODUÇÃO

Um dos principais objetivos da cirurgia primária do palato na fissura labiopalatina (FLP) é a reconstrução bem-sucedida da cinta muscular dos elevadores, de forma a propiciar um mecanismo velofaríngeo funcional para produção da fala adequada e a prevenção de complicações na orelha média, evitando comprometimento do crescimento facial¹. A ocorrência de fístulas oronasais residuais é um dos fatores indicativos do sucesso do reparo cirúrgico primário do palato²⁻⁴. A definição de fístula conforme reportado por Brosco⁵ (2017) é uma falha na cicatrização ou ruptura do reparo cirúrgico primário do palato e sua incidência varia entre 0% a > 60%⁶. Para prevenir e minimizar estas complicações cirúrgicas é importante entender melhor os fatores associados à ocorrência de fístulas.

A chamada “era da informação” caracteriza-se pela crescente expansão no volume de dados gerados e armazenados, fenômeno que também se reflete na área de saúde em geral, o que amplia possibilidade de obtenção de informações importantes no apoio ao processo decisório⁷. Os dados dos pacientes, bem como os resultados das cirurgias, ficam disponibilizados em seus prontuários, podendo ser usados como elementos para estudos clínicos. Porém, muitas vezes, o volume de dados gerados é tão grande que dificulta sua utilização e análise manual, demandando processos mais sofisticados como, por exemplo, os processos

automatizados, para a manipulação de tais dados. É exatamente neste contexto de superabundância de dados que surgiu a mineração de dados, como um processo sistemático, interativo e iterativo, de preparação e extração de conhecimento a partir de grandes bases de dados⁷.

OBJETIVO

Neste cenário, este trabalho tem por objetivo utilizar técnicas de mineração de dados para a descoberta de conhecimento sobre fatores associados à ocorrência de fístulas, após palatoplastia, em pacientes com fissura transforame incisivo unilateral (FTIU).

MÉTODOS

A investigação trata de uma pesquisa empírica, de natureza quantitativa e qualitativa. A amostra refere-se a um subconjunto de prontuários de pacientes com fissura labiopalatina participantes de um estudo clínico randomizado um estudo clínico randomizado (ECR) com FTIU⁸.

Dados sobre a ocorrência de fístulas foram obtidos para o total de 466 pacientes (bebês). Estes pacientes foram randomizados (usando-se um programa computacional) para receber diferentes protocolos de tratamento cirúrgico incluindo: 1) queiloplastia primária entre 3 e 6 meses de idade com a técnica de Millard (M) ou Spina (S); 2) palatoplastia precoce (9 a 12 meses) ou tardia (>12 meses); 3) palatoplastia primária com a técnica de von Langenbeck (VL) ou de Furlow (F); e 4) para um de quatro possíveis cirurgias (C1, C2, C3, C4).

De interesse para o presente estudo foram as informações sobre a ocorrência de fístula após a palatoplastia primária. Para determinação das classes

na mineração de dados usou-se a classificação de Spina⁹ (1972) agrupando-se os pacientes em dois grupos: SUCESSO (pacientes sem fístula ou com fístula na região pré-forame incisivo); INSUCESSO (pacientes com fístula na região pós-forame incisivo ou fístulas transforame). O forame incisivo demarca os limites dos palatos primário (parte central do lábio superior e pré-maxila) e secundário (palato duro e mole).

A Tabela 1 apresenta as informações identificadas nos prontuários dos pacientes para este estudo incluindo: técnica cirúrgica na queiloplastia e na palatoplastia; tempo da palatoplastia; idade na palatoplastia; cirurgião; uso de modificações na cirurgia como incisão relaxante e retalho de vômer; duração da palatoplastia em minutos; resultado quanto ocorrência de fístula (SUCESSO ou INSUCESSO); se ocorreu infecção na palatoplastia (no local) ou em outro local após a palatoplastia; se houve vômito ou tosse no pós-operatório da palatoplastia; se houve diagnóstico sintomático da disfunção velofaríngea, presença de hipernasalidade (registrada em conversa espontânea ou dirigida); resultados dos testes de emissão de ar nasal, de hipernasalidade e de hiponasalidade (observados durante repetição de 10 vocábulos). As variáveis de interesse estão listadas na coluna “nome do atributo”.

Tabela 1. Definição das variáveis (atributos) de interesse para este estudo

Variáveis (Nome do Atributo)	Categorias (Valores)
Técnica cirúrgica na queiloplastia	Millard, Spina
Tempo da palatoplastia	Precoce (9-12 m), Tardio (>12 m)
Idade na palatoplastia	Meses (m)
Técnica cirúrgica na palatoplastia	Furlow, Von Langenbeck
Cirurgião	C1, C2, C3, C4
Incisão relaxante	sem incisão, unilateral, bilateral
Retalho de Vômer	sim, não
Duração da palatoplastia	Minutos
Infecção na palatoplastia	não houve; no local da cirurgia, em outro local
Vômito no pós-operatório da palatoplastia	sim, não
Tosse no pós-operatório da palatoplastia	sim, não

Febre	sim, não
Sugestivo de disfunção velofaríngea	sim, não
Hipernasalidade	sim, não
Teste de emissão de ar	[1-10]
Teste de hipernasalidade	[1-10]
Teste de hiponasalidade	[1-10]
Ocorrência de fistula	SUCESSO, INSUCESSO

m=meses; C=cirurgião

No gerenciamento da FLP o SUCESSO do tratamento ocorre na ausência de fístula e ausência de alterações de fala. Para o presente estudo a ocorrência de fístula em região posterior ao forame incisivo e presença de disfunção velofaríngea foram interpretadas como indicativas de INSUCESSO do tratamento. A pergunta norteadora da mineração envolveu a verificação de quais fatores estariam associados à ocorrência ou não das fístulas após a palatoplastia primária. Buscou-se, portanto, identificar com este estudo se algumas das variáveis analisadas podem ser usadas como preditoras da ocorrência de fístula no palato.

Para o cômputo dos resultados do experimento foi utilizado o algoritmo C4.5 (J48) que gera árvores de decisão para encontrar a relação entre as características consideradas e os resultados das cirurgias bem como o algoritmo de associação a priori para a geração de regras. Nesta análise as variáveis de interesse foram tratadas como os atributos no software utilizado (WEKA). Considerando uma tarefa de mineração típica, o experimento foi dividido em quatro etapas: pré-processamento dos dados, extração de características, classificação e descrição de resultados. O procedimento foi realizado considerando-se como resultado primário a ocorrência de fistula após a palatoplastia. O pré-processamento foi realizado de maneira semiautomática. Os dados dos prontuários disponibilizados no formato de arquivo “.XLS” (planilha de Excel®) foram convertidos para o formato “.ARFF” com a

ajuda de um software open source (Excel2ArffConverter), usado pelo WEKA. Antes da conversão os atributos foram identificados conforme descrito na Tabela 1.

RESULTADOS

Somente os pacientes com dados completos foram selecionados para análise, considerando os parâmetros descritos na Tabela 1. Após descartar os pacientes com dados incompletos para qualquer uma das variáveis, um total de 222 instâncias de dados foram selecionadas para análise. Devido à possibilidade de enviesamento da base (somando ou subtraindo informações), optou-se por não se estimar os valores ausentes⁷. As informações sobre ocorrência de algum tipo de fístula foram identificadas nos prontuários de 222 (47,6%) dos 466 pacientes estudados e foram os dados destes pacientes que foram minerados para o presente artigo.

No grupo de 222 pacientes considerados para este estudo 98 (44,1%) indivíduos pertenciam ao sexo feminino e 124 (55,9%) ao sexo masculino e receberam a palatoplastia primária com idade média de 12,8 meses ($\sigma=3,2$). Neste grupo 114 (51,3%) recebeu o procedimento de Millard na queiloplastia primária enquanto 108 (48,7%) recebeu o procedimento de Spina. Um total de 112 pacientes (50,4%) recebeu a técnica de Furlow na palatoplastia primária enquanto 110 (49,6%) receberam von Langenbeck. Da amostra de 222 pacientes, 182 (81,9%) pertenciam ao grupo SUCESSO e 40 (18,1%) ao grupo INSUCESSO.

Por meio da construção de uma árvore de decisão, 37 regras foram geradas a partir do conjunto completo dos dados referentes aos pacientes. Entretanto, neste artigo optou-se por exibir apenas as 5 regras com maior cobertura de cada resultado final da cirurgia (SUCESSO ou INSUCESSO), isto é, com o maior número total de

instâncias atingindo o nó folha. Também são exibidas informações sobre a precisão da regra (probabilidade do resultado condicionado aos atributos).

A média de precisão das regras associadas ao SUCESSO cirúrgico é de 97,26% ($\sigma=4.59$). As cinco regras conseguem cobrir 141 instâncias, ou seja, juntas apresentam uma cobertura de cerca de 77,5%. Já para a classe de INSUCESSO, a média de precisão das regras associadas é de 84,32% ($\sigma=9.40$). A cobertura das cinco regras é de 62,5%. A regra com maior cobertura e precisão para predição de um bom resultado indica que os principais fatores envolvidos são: infecção (“ausência”), testes de hipernasalidade (“ ≤ 6 ”) e hiponalidade (“ > 9 ”) e a técnica cirúrgica (“Von Langenbeck”). Já para o INSUCESSO, segundo as duas regras com maior precisão e cobertura, os fatores envolvidos incluem: infecção (“ausência ou em outro local”), testes de hipernasalidade (“ > 6 ”), emissão de ar (“ > 9 ”) e febre (“sim”). As regras são exibidas na Tabela 2.

Tabela 2. Resultado da cirurgia

Número	Regra	Resultado (classe)	Cobertura	Precisão
1	Se "infecção=não houve" e "teste de hipernasalidade ≤ 6 " e "tosse=não" e "técnica cirúrgica=Von Langenbeck" e "teste de hiponasalidade > 9 "	SUCESSO	77	100%
2	Se "infecção=não houve" e "teste de hipernasalidade ≤ 6 " e "tosse=não" e "técnica cirúrgica=Furlow" e "amplitude da fissura=regular"	SUCESSO	33	96,9%
3	Se "infecção=não houve" e "teste de hipernasalidade ≤ 6 " e "tosse=não" e "técnica cirúrgica=Furlow" e "amplitude da fissura=ampla" e "cirurgião=C3"	SUCESSO	19	89,4%
4	Se "infecção=não houve" e "teste de hipernasalidade ≤ 6 " e "tosse=não" e "técnica cirúrgica=Furlow" e "amplitude da fissura=ampla" e "cirurgião=C2"	SUCESSO	7	100%
5	Se "infecção=não houve" e "teste de hipernasalidade ≤ 6 " e "tosse=não" e "técnica cirúrgica=Furlow" e "amplitude da fissura=ampla" e "cirurgião=C1" e "incisão relaxante=sem"	SUCESSO	5	100%
6	Se "infecção=não houve" e "teste de hipernasalidade > 6 " e "teste de emissão de ar > 9 " e "febre=sim"	INSUCESSO	6	83,3%

7	Se "infecção=ocorreu em outro local"	INSUCESSO	6	83,3%
8	Se "infecção=não houve" e "teste de hipernasalidade>6" e "teste de emissão de ar>9" e "febre=não" e "incisão relaxante=bilateral" e "vômito=não" e "cirurgião=C3"	INSUCESSO	5	80,0%
9	Se "infecção=não houve" e "teste de hipernasalidade<=6" e "tosse=não" e "técnica cirúrgica=Furlow" e "amplitude da fissura=ampla" e "cirurgião=C4" e "teste de emissão de ar>2"	INSUCESSO	4	100%
10	Se "infecção=não houve" e "teste de hipernasalidade>6" e "teste de emissão de ar>9" e "febre=não" e "incisão relaxante=sem"	INSUCESSO	4	75,0%

Quando se analisa o desempenho global do modelo (árvore de decisão gerada), observa-se que ele classifica corretamente 95,9% das instâncias e incorretamente apenas 4,1%. Considerando cada categoria individualmente, o modelo consegue acertar 90,0% dos casos em que ocorrem um resultado de INSUCESSO. Já para a outra classe, o modelo consegue acertar 97,3% dos casos em que ocorrem um resultado de SUCESSO.

As correlações encontradas usando o algoritmo de regras de associação (a priori) foram obtidas usando um suporte mínimo de 60% e uma confiança de 90%. O objetivo foi encontrar regras que fossem frequentes (alto suporte) na base de dados e com alto grau de confiança (diretamente relacionada à validade da regra). Foram encontradas quatro regras, com uma confiança média de 90,75% ($\sigma=0.5$) e um suporte médio de 69,45% ($\sigma=0.49$), que atendem os requisitos citados, conforme Tabela 3.

Considerando-se apenas o grupo de 40 pacientes do grupo de INSUCESSO os resultados mostram 6 regras encontradas com suporte mínimo de 67,5% e confiança mínima de 100% (Tabela 3). As regras apresentam um suporte médio de 72,08%. Os fatores associados ao grupo de SUCESSO envolvem ausência de tosse

e hipernasalidade e ocorrência de infecção. Pacientes do grupo de INSUCESSO também apresentaram ausência de tosse e infecção.

Tabela 3. Regras com alto valor de suporte e confiança

Características	Resultado	Suporte	Confiança
Ausência de tosse e infecção sem sugestivo de disfunção velofaríngea	SUCESSO	69,8%	91,0%
Ausência de tosse e infecção com hipernasalidade ausente	SUCESSO	69,8%	91,0%
Ausência de tosse e infecção sem sugestivo de disfunção velofaríngea com hipernasalidade ausente	SUCESSO	69,8%	91,0%
Ausência de tosse e infecção e sem febre	SUCESSO	68,4%	90,0%
Ausência de tosse	INSUCESSO	77,5%	100,0%
Ausência de infecção	INSUCESSO	77,5%	100,0%
Técnica cirúrgica de Furlow	INSUCESSO	72,5%	100,0%
Utilização de retalho de Vômer	INSUCESSO	70,0%	100,0%
Ausência de vômito	INSUCESSO	67,5%	100,0%
Ausência de tosse e de infecção	INSUCESSO	67,5%	100,0%

A Tabela 4 traz um resumo da relação entre a duração da palatoplastia e o resultado quanto à ocorrência de fístula. Observa-se que os tempos de cirurgia variam de 25 à 140 minutos.

Tabela 4. Relação entre duração da palatoplastia e classes (SUCESSO e INSUCESSO)

Duração: Minutos	N	Média	Desvio padrão	Mínimo	Máximo
Duração da palatoplastia – Todos os grupos	222	65,62	24,43	25	140
Duração da palatoplastia (grupo SUCESSO)	182	62,57	22,89	25	125
Duração da palatoplastia (grupo INSUCESSO)	40	79,5	26,62	25	140

DISCUSSÃO

Especificamente com relação às ocorrências de fístulas as regras encontradas com alto grau de precisão e cobertura podem trazer insights sobre quais fatores são determinantes para o sucesso ou insucesso da palatoplastia. A oportunidade de adotar a mineração sobre os dados de pacientes submetidos à

palatoplastia pode oportunizar melhor entendimento das especificidades que podem ocorrer com o grupo de pacientes, ampliando, assim, o conhecimento do profissional na identificação das condutas a serem adotadas.

Neste estudo específico, a visibilidade dada a alguns fatores (Tabela 1) dá possibilidades aos profissionais de saúde de, com a devida análise desse conjunto de descobertas, identificar padrões de associação de variáveis, as quais possam dar significado às ações diagnósticas e terapêuticas. Da mesma forma que em outros estudos prévios, esta investigação optou pela combinação de diferentes tipos de tarefas de mineração de dados para a realização do experimento ou identificação de padrões¹⁰⁻¹⁴.

Apesar da disponibilidade inicial de dados referentes à 466 pacientes, optou-se pelo uso de 222 (considerando apenas aqueles completos). Isso pode ter limitado as regras obtidas bem como não ter evidenciado outras associações dos fatores relativas aos resultados finais da palatoplastia. Tal decisão segue as diretrizes de outros trabalhos¹⁵. Para estudos futuros toda a base poderá ser usada, pois alguns algoritmos podem lidar com dados faltantes⁷. Outra limitação referente à base é o fato de as duas classes consideradas estarem desbalanceadas, entretanto, como elas refletem a situação real em que os resultados de SUCESSO são mais comuns que os INSUCESSOS, optou-se pela manutenção da proporção natural dos dados. Essa presença de classes majoritárias com frequência muito maior que as outras classes minoritárias, faz com que os algoritmos tenham uma tendência para responder bem para as classes majoritárias em detrimento das minoritárias. Como trabalhos futuros, o experimento poderá ser repetido usando técnicas de resampling aleatório (undersampling ou oversampling), isto é, uma reamostragem dos exemplos de treinamento de forma a gerar conjuntos balanceados¹⁶.

O fato de toda a base ter sido usada para o treinamento e teste, pode gerar um viés de ajuste aos dados. Qualquer método de aprendizado de máquina pode sofrer do que se chama de overfitting que é quando se produz um classificador que se ajusta bem aos dados de treinamento, mas não consegue generalizar o conhecimento aprendido, não obtendo um bom desempenho nos dados de teste. É como se o classificador decorasse os dados do treinamento. Entretanto, como o intuito deste experimento não é, necessariamente, automatizar o processo de classificação, mas sim, gerar regras que possam ser avaliadas por humanos, gerando possíveis insights, optou-se por esta abordagem.

A análise da Tabela 2 indica que os resultados de SUCESSO estão associados à ausência de infecção e tosse e, além disso, os pacientes apresentaram teste de hipernasalidade abaixo ou igual a 6 (numa escala que vai até 10). No caso das fissuras amplas associadas à técnica cirúrgica de Furlow, além dos fatores ressaltados, há influência do fator cirurgião para o resultado final.

No caso dos resultados de INSUCESSO, a presença de infecção parece ser um fator importante, entretanto, não é decisivo. Devido à similaridade entre as regras 9, 3 e 4 (Tabela 2) o fator decisivo para obtenção de um resultado de INSUCESSO está atrelado ao cirurgião. Sob as mesmas condições os cirurgiões C2 e C3 obtiveram resultados de SUCESSO, entretanto, o cirurgião C4 obteve SUCESSO em apenas em 50% das cirurgias, o que pode indicar a influência do fator cirurgião. Valores de testes de hipernasalidade superiores à 6 são indicativos de um possível INSUCESSO.

O algoritmo apriori apresentou como resultado regras de associação que, do mesmo modo que as regras da árvore de decisão, deverão ser avaliadas por um profissional, de forma a validá-las frente à realidade. Por conta das medidas de

suporte e confiança apenas algumas variáveis foram consideradas nas regras. Durante os experimentos realizados com o apriori, percebeu-se que não é suficiente aumentar o número de regras ou reduzir a medida de suporte mínimo, visando obter mais informação. Isto porque o algoritmo começa a induzir regras redundantes que não acrescentam informação. Uma estratégia adotada foi, portanto, somente priorizar as regras com alto suporte e confiança. O algoritmo apriori não lida com atributos quantitativos, somente com os categóricos, o que exige a exclusão de alguns atributos ou mesmo a transformação deles para dados não numéricos (processo de discretização), essa estratégia foi utilizada em alguns processamentos realizados neste trabalho. Assim, para evitar essa limitação em trabalhos futuros outros algoritmos poderão ser experimentados tais como AprioriTid, SETM, AprioriHybrid¹⁷.

Com relação à tarefa de associação, a análise da Tabela 3 indica que, de um modo geral, os fatores combinados como ausência de infecção e de febre, ausência de hipernasalidade e paciente sem sugestivo de disfunção velofaríngea apresentam SUCESSO pós palatoplastia primária. Com relação aos procedimentos cirúrgicos há indícios de que a utilização da técnica de Furlow e de retalho de Vômer são frequentes no grupo de INSUCESSO. Já sintomas como ausência de tosse, vômito ou infecção, isoladamente, não podem ser usados como parâmetros para descartar um possível INSUCESSO. A análise da Tabela 4 mostra que uma palatoplastia do grupo dos pacientes que tiveram um resultado de INSUCESSO dura em média 79,5 minutos; já para o grupo de pacientes com resultados de SUCESSO a média cai para 62,57 minutos. Há indícios, portanto, que cirurgias mais demoradas causam piores resultados.

Reconhece-se, finalmente, que este estudo oferece somente uma perspectiva pontual da realidade, por meio das análises preliminares de técnicas de mineração de dados na base considerada, já que revela apenas alguns fatores associados aos resultados da palatoplastia do ponto de vista dos algoritmos de aprendizados de máquina, havendo a necessidade de validação por profissionais da área de saúde.

CONCLUSÃO

A análise dos dados revelou que a ausência de alguns sintomas (febre, tosse, infecção) bem como características associadas à cirurgia em si (cirurgião, técnica, retalho de vômer) e ao paciente (hipernasalidade, sugestivo de disfunção velofaríngea), podem ajudar a predizer o sucesso ou insucesso da palatoplastia.

REFERÊNCIAS

1. Dutka, JCR; Pegoraro-Krook MI. Avaliação e tratamento das disfunções velofaríngeas. In: Marchesan IQ, editor. Tratado de especialidades em Fonoaudiologia. São Paulo: Grupo Gen-Editora Roca Ltda; 2014. p. 363–68.
2. Passos VAB, Carrara CFC, Dalben GS, Costa B, Gomide MR. Prevalence, Cause, and Location of Palatal Fistula in Operated Complete Unilateral Cleft Lip and Palate: Retrospective Study. *Cleft Palate-Craniofacial J.* 2014 51(2):158–64.
3. Deshpande GS, Campbell A, Jagtap R, Restrepo C, Dobie H, Keenan HT, et al. Early Complications After Cleft Palate Repair. *J Craniofac Surg.* 2014 25(5):1614–8.

4. Brosco TVS. Fístula de palato após reparo da fissura labiopalatina em um estudo clínico randomizado [Tese de doutorado]. São Paulo: Hospital de Reabilitação de Anomalias Craniofaciais, Universidade de São Paulo; 2017. 168p.
5. Brosco TVS, Prearo GA, Silva HLA, Dutka JCR. Brosco-Dutka classification system for palate fistulas. *Rev Bras Cir Plástica – Brazilian J Plast Sugery*. 2021;36(2):164–72.
6. Muzaffar AR, Byrd SH, Rohrich RJ, Johns DF, LeBlanc D, Beran SJ, et al. Incidence of Cleft Palate Fistula: An Institutional Experience with Two-Stage Palatal Repair. *Plast Reconstr Surg*. 2001;108(6):1515–8.
7. Castro, LN; Ferrari DG. Introdução à mineração de dados: Conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva; 2017.
8. Dutka J de CR. Estudo clínico randomizado - Projeto Florida (ECR-PF-ementa fase 2): função velofaríngea para a fala e estudo do crescimento da face e dos arcos dentários após a palatoplastia primária. Projeto em andamento e com aprovação Ética do CEP/CONEP desde 02/09/2016. Pesquisador Responsável: Jeniffer de Cássia Rillo Dutka. CAAE: 57727416.9.0000.5441. Instituição Proponente: Hospital de Reabilitação de Anomalias Craniofaciais da USP
9. Spina V et al. Classificação das fissuras lábio-palatais: sugestão de modificação. *Rev Hosp Clín Fac Med São Paulo*. 1972;27:5–6.
10. West D, Mangiameli P, Rampal R, West V. Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *Eur J Oper Res*. 2005;162(2):532–51.

11. Steiner MTA, Soma NY, Shimizu T, Nievola JC, Steiner Neto PJ. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. *Gestão & Produção*. 2006;13(2):325–37.
12. Junior ACB. Utilização de técnicas de data mining na detecção de outliers em auxílio à auditoria operacional com um estudo de caso com dados do sistema de informações hospitalares. Universidade Federal do Rio de Janeiro; 2009.
13. Kuretzki, CH. Técnicas de mineração de dados aplicadas em bases de dados da saúde a partir de protocolos eletrônicos. [Dissertação de mestrado]. Curitiba: Universidade Federal do Paraná; 2009. 98p.
14. Martínez GRS, Bermúdez YVC. Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial. *Rev Tecnura*. 2012; 16(33):35.
15. Carvalho DR, Moser AD, Silva VA da, Dallagassa MR. Mineração de Dados aplicada à fisioterapia. *Fisioter em Mov*. 2012;(3):595–605.
16. Goldschmidt, Ronaldo; Passos, Emmanuel; Bezerra E. *Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações*. 2nd ed. Rio de Janeiro: Elsevier; 2015.
17. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. *ACM SIGMOD Rec*. 1993 Jun;22(2):207–16.

6

Conclusão Geral

6 CONCLUSÃO GERAL

O Hospital de Reabilitação de Anomalias Craniofaciais da Universidade de São Paulo (HRAC-USP), Bauru, SP-Brasil, em colaboração com o Centro Craniofacial da Universidade da Florida (UFCFC), estabeleceu um estudo clínico randomizado (ECR) denominado Projeto Florida (PF), que vem investigando, ao longo de mais de 20 anos, os resultados do tratamento cirúrgico primário para corrigir a fissura transforame unilateral (WILLIAMS et al., 1998; WILLIAMS et al., 2011). Desde 1996 os dados obtidos no ECR-PF são registrados em prontuários do paciente e quando digitados são armazenados em uma base de dados mantida em um servidor destinado unicamente à gestão desses dados (sistema legado). O acervo permite o estudo de dados anonimizados e possibilita a obtenção de tabelas eletrônicas que são usadas em diversos projetos envolvendo subconjuntos dos dados do ECR-PF. A mineração desses dados, por sua vez, pode colaborar para modernizar a análise e a interpretação dos achados, contribuindo por meio do viés computacional para uma pesquisa mais detalhada dos achados clínicos documentados ao longo dos anos do ECR-PF (RIBEIRO *et al.*, 1998; DE ESPINDOLA; MAJDENBAUM; AUDY, 2004; PINTO e BRAGA, 2005; ALLIEVI, 2015). As análises de dados automatizadas com aplicações de técnicas de inteligência artificial são importantes para gerar hipóteses, compreender relações e encontrar padrões inesperados em bases de dados hospitalares, complementando os demais estudos realizados.

Além das contribuições pontuais desta pesquisa evidenciadas nos artigos que compõem esta tese, podem ser evidenciadas outras contribuições potenciais deste trabalho:

- Foi possível ampliar as parcerias entre as áreas de ciência da computação e ciências da saúde, sobretudo no escopo das fissuras labiopalatinas, buscando-se otimizar processos relacionados às práticas dessas ciências, por meio das tecnologias da informação e comunicação, fomentando projetos multidisciplinares;
- O presente trabalho de mineração de dados serve como fonte de consulta sobre o processo de descoberta de conhecimento em

bases de dados da saúde, sobretudo, relacionadas às fissuras labiopalatinas, uma vez que o assunto é relativamente novo na prática clínica e não existe grande disponibilidade de bibliografia, principalmente, no contexto brasileiro;

- O desenvolvimento de pesquisa envolvendo assuntos de diferentes áreas como saúde, computação, estatística e descoberta de conhecimento oferece perspectivas para o desenvolvimento de novos projetos, sob uma ótica multidisciplinar e interprofissional;
- A utilização de ferramentas distintas e algoritmos de mineração de dados e aprendizado de máquina como os usados nos dois estudos apresentados, incluindo, ambiente de mineração como o software WEKA ou linguagens de programação voltadas à análise de dados como a Python, favorece a implementação de novos modelos de análise junto à equipe de pesquisadores do ECR-PF, uma vez que permite a automatização e aprofundamento das análises de dados;
- De forma pontual, as análises conduzidas revelaram alguns padrões: a ausência de algumas complicações pós-operatórias (febre, vômito, tosse, infecção) bem como características associadas à cirurgia em si (cirurgião, duração, técnicas cirúrgicas, retalho de Vômer, incisão relaxante) e ao paciente (hipernasalidade, sugestivo de disfunção velofaríngea, idade na palatoplastia), podem ajudar a prever o sucesso ou insucesso da palatoplastia, com relação à ocorrência de fístulas que são consideradas complicações. Tais resultados e padrões de conhecimento descobertos poderão ser utilizados em ações práticas e futuros estudos clínicos, salientando a importância da aproximação entre profissionais de informática e saúde tanto no ambiente acadêmico quanto na implementação de práticas clínicas baseadas em evidências científicas.

Referências

REFERÊNCIAS

AGRAWAL, Rakesh; PSAILA, Giuseppe. Active Data Mining. **KDD**, [s. l.], p. 3-8, 1995.

ALLIEVI, Odalis. **Reflexões sobre manutenção de sistemas legados**. [S. l.], 17 abr. 2207. Disponível em: <https://imasters.com.br/desenvolvimento/reflexoes-sobre-manutencao-de-sistemas-legados>. Acesso em: 14 set. 2021.

ARMAÑANZAS, Rubén; BIELZA, Concha; CHAUDHURI, Kallol Ray; MARTINEZ-MARTIN, Pablo; LARRAÑAGA, Pedro. Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach. **Artificial intelligence in medicine (Elsevier)**, [s. l.], v. 58, n. 3, p. 195-202, 2013.

BAGUI, Subhash C; BAGUI, Sikha; PAL, Kuhu; PAL, Nikhil R. Breast cancer detection using rank nearest neighbor classification rules. **Pattern recognition**, [s. l.], v. 36, n. 1, p. 25-34, 2003.

BARELLA, Victor Hugo. **Imbalanced classification tasks: measuring data complexity and recommending techniques**. Orientador: André Carlos Ponce de Leon Ferreira de Carvalho. 2021. Tese (Doutorado em Ciências) - Instituto de Ciências Matemáticas e de Computação, São Carlos, 2021. DOI <https://doi.org/10.11606/T.55.2021.tde-26042021-140437>. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-26042021-140437/publico/VictorHugoBarella_revisada.pdf. Acesso em: 15 jun. 2021.

BHATTACHARYYA, Dhruva K; HAZARIKA, Syamanta M. **Networks, data mining, and artificial intelligence: trends and future directions**. 1. ed. [S. l.]: Narosa Pub House, 2006. ISBN 978-8173197550.

CHEN, Daniel Y. **Análise de Dados com Python e Pandas**. 1. ed. São Paulo: Novatec, 2018. ISBN 978-8575226995.

DATE, Christopher J. **Introdução a sistemas de bancos de dados**. 1. ed. São Paulo: GEN LTC, 2004. 896 p. ISBN 978-8535212730.

DE ESPINDOLA, Rodrigo Santos; MAJDENBAUM, Azriel; AUDY, Jorge Luis Nicolas. Uma Análise Crítica dos Desafios para Engenharia de Requisitos em Manutenção de Software. **VII Workshop on Requirements Engineering**, Tandil-Argentina, p. 226-238, 2004.

DHILLON, Harnoor; CHAUDHARI, Prabhat Kumar; DHINGRA, Kunaal; KUO, Rong-Fu; SOKHI, Ramandeep Kaur; ALAM, Mohammad Khursheed; AHMAD, Shandar. Current Applications of Artificial Intelligence in Cleft Care: A Scoping Review. **Frontiers in Medicine**, v. 8, 2021.

EL-HALEES, Alaa M; ALMADHOUN, Mohamed D. Different mining techniques for health care data case study of urine analysis test. **International Journal of**

Biomedical Data Mining, [s. l.], v. 6, n. 2, 2017. DOI 10.4172/2090-4924.1000129. Disponível em: https://www.researchgate.net/publication/322660243_Different_Mining_Techniques_for_Health_Care_Data_Case_Study_of_Urine_Analysis_Test. Acesso em: 6 abr. 2021

FAYYAD, Usama M. Mining databases: Towards algorithms for knowledge discovery. **Bulletin of the Technical Committee on Data Engineering**, [s. l.], v. 21, n. 1, p. 39-48, 1998.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, [s. l.], v. 17, n. 3, p. 37-37, 1996. DOI <https://doi.org/10.1609/aimag.v17i3.1230>. Disponível em: <https://ojs.aaai.org//index.php/aimagazine/article/view/1230>. Acesso em: 13 out. 2020.

FEI, Sheng-wei. Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine. **Expert Systems with Applications**, [s. l.], v. 37, n. 10, p. 6748-6752, 2010. DOI <https://doi.org/10.1016/j.eswa.2010.02.126>. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S095741741000165X>. Acesso em: 12 fev. 2019.

FERRARI, Daniel Gomes; SILVA, Leandro Nunes de Castro. **Introdução a mineração de dados**. 1. ed. São Paulo: Saraiva Uni, 2016. 376 p. ISBN 978-8547200985.

GAMA, João; CARVALHO, André Ponce de Leon; FACELI, Katti; LORENA, Ana Carolina; OLIVEIRA, Márcia. **Extração de Conhecimento de Dados: Data Mining**. 3. ed. [S. l.]: Sílabo, 2017. 436 p. ISBN 978-9726189145.

GARCÍA-LAENCINA, Pedro J; ABREU, Pedro Henriques; ABREU, Miguel Henriques; AFONOSO, Noémia. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. **Computers in biology and medicine**, [s. l.], v. 59, p. 125-133, 2015. DOI <https://doi.org/10.1016/j.compbio.2015.02.006>. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0010482515000554?via%3Dihub>. Acesso em: 14 set. 2021.

GENKIN, Alexander; LEWIS, David D; MADIGAN, David. Large-scale Bayesian logistic regression for text categorization. **Technometrics**, [s. l.], v. 49, n. 3, p. 291-304, 2007. DOI 10.1198/004017007000000245. Disponível em: https://www.researchgate.net/publication/355683518_Robust_model-based_estimation_for_binary_outcomes_in_genomics_studies. Acesso em: 5 out. 2021.

GUILLIAMS, Evi. **Improving patient safety using data mining techniques and ICD-9 codes**. Research group Policy Management. Diepenbeek-Bélgica: Research group Policy Management – Patient safety - Hasselt University, 2008. Disponível em: <https://silo.tips/download/improving-patient-safety-using-data-mining-techniques-and-icd-9-codes#>. Acesso em: 13 jul. 2021.

HANSEN, Peter Wæde; CLEMMENSEN, Line; SEHESTED, Thomas S. G. Identifying drug--drug interactions by data mining: a pilot study of warfarin-associated drug interactions. **Circulation Cardiovascular Quality and Outcomes**, [s. l.], v. 9, n. 6, p. 621-628, 2016. DOI 10.1161/CIRCOUTCOMES.116.003055. Disponível em: https://www.researchgate.net/publication/309874089_Identifying_Drug-Drug_Interactions_by_Data_Mining_A_Pilot_Study_of_Warfarin-Associated_Drug_Interactions. Acesso em: 4 ago. 2021.

HAUGHOM, John. Knowledge Management in Healthcare: It's more important than you realize. *In: Health Catalyst*. [S. l.], 2014. Disponível em: <https://www.healthcatalyst.com/enable-knowledge-management-in-healthcare>. Acesso em: 20 dez. 2019.

HEUSER, Carlos Alberto. **Projeto de Banco de Dados**. 6. ed. São Paulo: Bookman, 2008. 282 p. ISBN 978-8577803828.

HILL, Thomas; LEWICKI, Paul. **Statistics: Methods and Applications**. 1. ed. [S. l.]: StatSoft, 2005. ISBN 978-1884233593.

JEN, Chih-Hung; WANG, Chien-Chih; JIANG, Bernard C; CHU, Yan-Hua; CHEN, Ming-Shu. Application of classification techniques on development an early-warning system for chronic illnesses. **Expert Systems with Applications**, [s. l.], v. 39, n. 10, p. 8852-8858, 2012.

JOTHI, Neesha; HUSAIN, Wahidah. Data mining in healthcare: a review. **Procedia computer science**, [s. l.], v. 72, p. 306-313, 2015. DOI <https://doi.org/10.1016/j.procs.2015.12.145>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050915036066/pdf?md5=cb9681638e84466311d3f9ee5d51da22&pid=1-s2.0-S1877050915036066-main.pdf>. Acesso em: 15 jun. 2021.

JOUDAKI, Hossein; RASHIDIAN, Arash; MINAEI-BIDGOLI, Behrouz; MAHMOODI, Mahmood; GERAILI, Bijan; NASIRI, Mahdi; MOHAMMAD, Arab. Using data mining to detect health care fraud and abuse: a review of literature. **Global journal of health science**, [s. l.], v. 7, n. 1, p. 194-202, 2015. DOI 10.5539/gjhs.v7n1p194. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4796421/>. Acesso em: 20 maio 2020.

KANG, Seokho; KANG, Pilsung; KO, Taehoon; CHO, Sungzoon; RHEE, Su-jin; YU, Kyung-Sang. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. **Expert Systems with Applications**, [s. l.], v. 42, n. 9, p. 4265-4273, 2015.

KANTARDZIC, Mehmed. **Data mining: concepts, models, methods, and algorithms**. 3. ed. [S. l.]: Wiley-IEEE Press, 2019. 672 p. ISBN 978-1119516040.

KAREGAR, M; ISAZADEH, Ayaz; FARTASH, Farzaneh; SADARI, Taha; NAVIN, A Habibizad. Data-mining by probability-based patterns. **ITI 2008-30th International Conference on Information Technology Interfaces**, [s. l.], p. 353-360, 2008.

- KEIM, Daniel A. Information Visualization and Visual Data Mining. **IEEE transactions on Visualization and Computer Graphics**, [s. l.], v. 8, n. 1, p. 1-8, 2002.
- KORTH, Henry F; SILBERSCHATZ, Abraham; SUDARSHAN, S. **Sistema de Banco de Dados**. 5. ed. São Paulo: Elsevier, 2006. 808 p. ISBN 978-8535211078.
- KUMAMARU, Hiraku; LEE, Moa P; CHOUDHRY, Niteesh K; DONG, Yaa-Hui; KRUMME, Alexis A; KHAN, Nazleen; BRILL, Gregory; KOHSAKA, Shun; MIYATA, Hiroaki; SCHNEEWEISS, Sebastian. Using previous medication adherence to predict future adherence. **Journal of managed care & specialty pharmacy**, [s. l.], v. 24, n. 11, p. 1146-1155, 2018. DOI <https://doi.org/10.18553/jmcp.2018.24.11.1146>. Disponível em: <https://www.jmcp.org/doi/full/10.18553/jmcp.2018.24.11.1146>. Acesso em: 4 mar. 2020.
- LIAO, Shu-Hsien; CHU, Pei-Hui; HSIAO, Pei-Yuan. Data mining techniques and applications: A decade review from 2000 to 2011. **Expert systems with applications**, [s. l.], v. 39, n. 12, p. 11303-11311, 2012.
- LIU, Bo; XIAO, Yanshan; CAO, Longbing; HAO, Zhifeng; DENG, Feiqi. Svdd-based outlier detection on uncertain data. **Knowledge and information systems**, [s. l.], v. 34, n. 3, p. 597-618, 2013.
- LUGER, George F. **Inteligência Artificial: Estruturas e estratégias para a solução de problemas complexos**. 4. ed. São Paulo: Bookman, 2004. 774 p. ISBN 978-8536303963.
- MAMIYA, Hiroshi; SCHWARTZMAN, Kevin; VERMA, Aman; JAUVIN, Christian; BEHR, Marcel; BUCKERIDGE, David. Towards probabilistic decision support in public health practice: Predicting recent transmission of tuberculosis from patient attributes. **Journal of biomedical informatics**, [s. l.], v. 53, p. 237-242, 2015.
- MARIN, Heimar de Fátima; WHITAKER, Iveth Yamagushi; SOUZA, Maria de Jesus Castro Souza; SAPAROLLI, Eliana Campos Leite. Informática em saúde: uma nova proposta aos profissionais de enfermagem. **Acta Paulista de Enfermagem**, [s. l.], v. 3, n. 4, p. 160-160, 1990. Disponível em: https://acta-ape.org/wp-content/uploads/articles_xml/1982-0194-ape-S0103-21001990000300079/1982-0194-ape-S0103-21001990000300079.pdf. Acesso em: 29 out. 2019.
- MARTINEZ, Julia C; KING, Martha P; CAUCHI, Richard. Improving the health care system: seven state strategies. **National Conference of State Legislatures**, [s. l.], p. 1-28, 2016. Disponível em: <https://www.ncsl.org/Portals/1/Documents/Health/ImprovingHealthSystemsBrief16.pdf>. Acesso em: 20 jan. 2020.
- MINOR, Lloyd B. Harnessing the power of data in health. **Stanford Medicine 2017 Health Trends Report**, [s. l.], 2017. Disponível em: <https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf>. Acesso em: 23 set. 2021.

PINTO, Herbert Laroça Mendes; BRAGA, José Luís. Sistemas Legados e as Novas Tecnologias: técnicas de integração e estudo de caso. **Informática Pública**, Belo Horizonte, v. 7, n. 1, p. 48-69, 2005. Disponível em: http://pbh.gov.br/informaticapublica/ANO7_N1_PDF/IP7N1_mendespinto.pdf. Acesso em: 14 dez. 2020.

PRESSMAN, Roger S. **Engenharia de Software**. 6. ed. São Paulo: Mcgraw-Hill Interamericana, 2006. ISBN 978-8586804571.

RÉMY, Nfongourain Mougoutou; MARTIAL, Tekinzang Tedondjio; CLÉMENTIN, Tayou Djamegni. The prediction of good physicians for prospective diagnosis using data mining. **Informatics in medicine unlocked**, [s. l.], v. 12, p. 120-127, 2018. DOI <https://doi.org/10.1016/j.imu.2018.07.005>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352914818300467>. Acesso em: 9 mar. 2021.

RIBEIRO, Nuno Palmeiro; DE ALMEIDA, Alberto Bigotte; ABREU, Fernando Brito; SOUSA, Pedro. Levantamento de Modelos de Dados em Sistemas Legados. **Sistemas de Informação**, Lisboa-Portugal, n. 9, p. 19-28, 1998. Disponível em: https://www.researchgate.net/profile/Pedro-Sousa-22/publication/230579324_Levantamento_de_Modelos_de_Dados_em_Sistemas_Legados/links/54901ade0cf225bf66a81613/Levantamento-de-Modelos-de-Dados-em-Sistemas-Legados.pdf. Acesso em: 16 out. 2019.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial**. 2. ed. São Paulo: Elsevier, 2004. 1040 p. ISBN 978-8535211771.

SABBATINI, Renato M.E. **Evolução dos Recursos de Informática na Saúde**. [S. l.], 2004. Disponível em: <https://www.sabbatini.com/renato/papers/EvolucaoRecursosInformaticaSaude.pdf>. Acesso em: 23 fev. 2021.

ŞAHAN, Seral; POLAT, Kemal; KODAZ, Halife; GÜNEŞ, Salih. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. **Computers in Biology and Medicine**, [s. l.], v. 37, n. 3, p. 415-423, 2007.

SAMANTA, Biswanath; BIRD, Geoffrey L; KUIJPERS, Marijn; ZIMMERMAN, Robert A; JARVIK, Gail P; WERNOVSKY, Gil; CLANCY, Robert R; LICHT, Daniel J; GAYNOR, J William; NATARAJ, Chandrasekhar. Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. **Artificial intelligence in medicine**, [s. l.], v. 46, n. 3, p. 201-215, 2009.

SHARMA, Neha; OM, Hari. Data mining models for predicting oral cancer survivability. **Network Modeling Analysis in Health Informatics and Bioinformatics**, [s. l.], v. 2, n. 4, p. 285-295, 2013. DOI <https://doi.org/10.1007/s13721-013-0045-7>. Disponível em: <https://link.springer.com/article/10.1007/s13721-013-0045-7>. Acesso em: 2 mar. 2021.

SHAW, William. Global strategies to reduce the health care burden of craniofacial anomalies: report of WHO meetings on international collaborative research on craniofacial anomalies. **The Cleft palate-craniofacial journal**, [s. l.], v. 41, n. 3, p. 238-243, 2004.

SU, Chao-Ton; WANG, Pa-Chun; CHEN, Yan-Cheng; CHEN, Li-Fei. Data mining techniques for assisting the diagnosis of pressure ulcer development in surgical patients. **Journal of medical systems**, [s. l.], v. 36, n. 4, p. 2387-2399, 2012.

THOMPSON, Vetta L Sanders; LANDER, Sean; XU, Shuyu; SHYU, Chi-Ren. Identifying key variables in African American adherence to colorectal cancer screening : the application of data mining. **BMC Public Health**, [s. l.], v. 14, n. 1, p. 1-10, 2014. DOI <https://doi.org/10.1186/1471-2458-14-1173>. Disponível em: <https://bmcpublihealth.biomedcentral.com/articles/10.1186/1471-2458-14-1173>. Acesso em: 7 abr. 2020.

TRAINA, Agma J M; TRAINA JR, Caetano; BOTELHO, Elisângela; BARIONI, Maria Camila Nardini; BUENO, Renato. Visualização de dados em sistemas de bancos de dados relacionais. **Simpósio Brasileiro de Banco de Dados**, Rio de Janeiro, p. 95-109, 2001.

VELOSO, Rui; PORTELA, Filipe; SANTOS, Manuel Filipe; SILVA, Álvaro; RUA, Fernando; ABELHA, António; MACHADO, José. A Clustering Approach for Predicting Readmissions in Intensive Medicine. **Procedia Technology**, [s. l.], v. 16, p. 1307-1316, 2014. DOI <https://doi.org/10.1016/j.protcy.2014.10.147>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2212017314003740>. Acesso em: 5 jan. 2021.

VISELTEAR, Arthur J. C.-EA Winslow and the early years of public health at Yale, 1915-1925. **The Yale journal of biology and medicine**, [s. l.], v. 55, n. 2, p. 137-151, 1982. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2596005/>. Acesso em: 2 dez. 2019.

VLAHOS, George E; FERRATT, Thomas W; KNOEPFLE, George. The use of computer-based information systems by German managers to support decision making. **Information & Management**, [s. l.], v. 41, n. 6, p. 763-779, 2004.

WANG, Kung-Jeng; MAKOND, Bunjira; WANG, Kung-Min. An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. **BMC medical informatics and decision making**, [s. l.], v. 13, n. 1, p. 1-14, 2013. DOI <https://doi.org/10.1186/1472-6947-13-124>. Disponível em: <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-124>. Acesso em: 30 mar. 2021.

WICKRAMASINGHE, Nilmini; SHARMA, Sushil; GUPTA, Jatinder. Knowledge Management in Healthcare. *In: MEDICAL Informatics: Concepts, Methodologies, Tools, and Applications*. 1. ed. [S. l.]: IGI Global, 2008. cap. 16, p. 186-197. ISBN 9781605660516.

WITTEN, Ian H; FRANK, Eibe; HALL, Mark A; PAL, Christopher. **Data Mining: Practical Machine Learning Tools and Techniques**. 4. ed. [S. l.]: Morgan Kaufmann, 2016. 654 p. ISBN 978-0128042915.

WORLD internet usage and population statistics 2021 year-q1 estimates. *In*: **INTERNETWORLDSTATS.COM**. [S. l.]: Miniwatts Marketing Group, 2021. Disponível em: <https://www.internetworldstats.com/stats.htm>. Acesso em: 15 set. 2021.

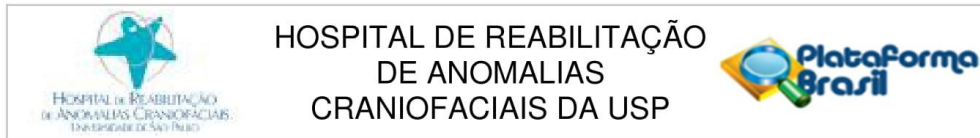
YANG, Ji-Jiang; LI, Jianqiang; MULDER, Jacob; WANG, Yongcai; CHEN, Shi; WU, Hong; WANG, Qing; PAN, Hui. Emerging information technologies for enhanced healthcare. **Computers in industry**, [s. l.], v. 69, p. 3-11, 2015.

ZHENG, Bichen; YOON, Sang Won; LAM, Sarah S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. **Expert Systems with Applications**, [s. l.], v. 41, n. 4, p. 1476-1482, 2014.

ZOLBANIN, Hamed Majidi; DELEN, Dursun; ZADEH, Amir Hassan. Predicting overall survivability in comorbidity of cancers: A data mining approach. **Decision Support Systems**, [s. l.], v. 74, p. 150-161, 2015.

Anexos

ANEXO A – PARECER DO COMITÊ DE ÉTICA



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: APLICAÇÃO DA MINERAÇÃO DE DADOS PARA APOIO À DESCOBERTA DE CONHECIMENTO NO CONTEXTO DO ESTUDO CLÍNICO RANDOMIZADO-PROJETO FLORIDA

Pesquisador: Patrick Pedreira Silva

Área Temática:

Versão: 1

CAAE: 59951416.6.0000.5441

Instituição Proponente: Hospital de Reabilitação de Anomalias Craniofaciais da USP

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.753.467

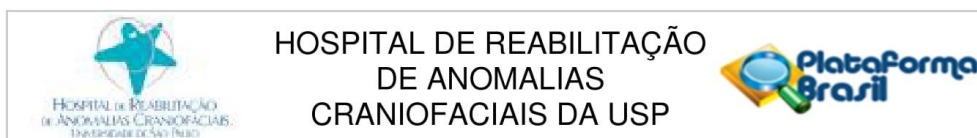
Apresentação do Projeto:

Trata-se de um projeto com a finalidade de tese de doutorado de autoria do aluno Patrick Pedreira Silva, sob orientação da Profª Drª Jeniffer de Cássia Rillo Dutka. A proposta é desenvolver um sistema para compilação de dados em formato eletrônico possibilitando modernizar a sua análise, aplicando técnicas de mineração e, ao mesmo tempo, permitindo a continuidade à informatização dos achados clínicos coletados no HRAC/USP. As fontes de dados que farão parte do estudo são referentes ao Estudo Clínico Randomizado-Projeto Florida (ECR-PF) e serão disponibilizados a partir de acesso ao servidor do ECR-PF, com aprovação da investigadora principal do ECR no Brasil, Profa Dra Maria Inês Pegoraro-Krook e da coordenadora administrativa do projeto (orientadora deste trabalho) Profa Dra Jeniffer de Cássia Rillo Dutka. Os dados de interesse foram coletados no período de 1996 a 2017 a partir dos protocolos clínicos incluídos nos prontuários dos 475 pacientes que completaram a fase 1 do ECR. Segundo os pesquisadores, essas técnicas aplicadas aos dados do Projeto Florida podem contribuir, decisivamente, nos estudos clínicos ligados às fissuras labiopalatinas.

Objetivo da Pesquisa:

O pesquisador definiu como objetivo primário do estudo "implantar uma base de dados eletrônica para uso de técnicas de mineração de dados no contexto das fissuras labiopalatinas,

Endereço: SILVIO MARCHIONE 3-20
Bairro: VILA NOVA CIDADE UNIVERSITARIA **CEP:** 17.012-900
UF: SP **Município:** BAURU
Telefone: (14)3235-8421 **Fax:** (14)3234-7818 **E-mail:** uep_projeto@centrinho.usp.br



Continuação do Parecer: 1.753.467

contemplando mais de duas décadas (1996 a 2017) de documentação dos resultados do Estudo Clínico Randomizado-Projeto Florida".

Como objetivos secundários foram definidos:

- Implantar a infraestrutura (hardware e software) para acomodar uma base de dados;
- Migrar a base de dados do sistema legado (Data Entry) para uma base de dados relacional;
- Estabelecer uma interface amigável para captura eletrônica dos dados, permitindo que a equipe do Projeto Florida dê continuidade à digitação dos achados clínicos que desde 2008 vêm sendo coletados pela equipe, mas não estão sendo capturados eletronicamente;
- Levantar, avaliar e aplicar ferramentas que permitam a extração de conhecimento por meio das técnicas de Mineração de Dados (Data Mining) à base de dados relacional;
- Avaliar a viabilidade e eficiência das técnicas de mineração de dados no contexto do ECR-PF, no que se refere à descoberta de conhecimento, visando identificar fatores de risco e fatores de proteção que possam prever os resultados de fala, função velofaríngea, crescimento e estética facial, após a correção primária da fissura labiopalatina.

Avaliação dos Riscos e Benefícios:

Segundo o pesquisador, o estudo "não envolverá acesso direto a seres humanos. Fontes secundárias de dados (base de dados do ECR-PF) serão acessadas e manuseadas. O acesso aos dados do ECR-PF será feito respeitando-se a garantia do resguardo das informações do paciente e da instituição sem riscos envolvidos".

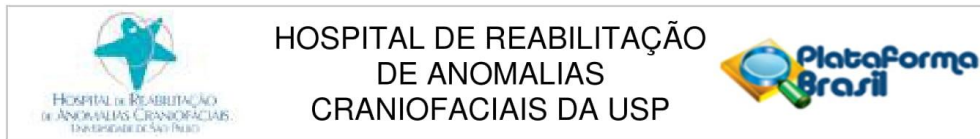
O pesquisador definiu como benefício deste estudo "a implementação do novo sistema de gerenciamento dos dados do Projeto Florida com uma nova interface de acesso, juntamente com migração da antiga base de dados e o uso de algoritmos de mineração de dados (data mining), representa uma grande contribuição para a consolidação do ECR-PF

e interpretação e publicação de seus achados. Este novo sistema poderá ser usado no HRAC para o gerenciamento de dados que permitam um monitoramento sistemático dos resultados do tratamento da FLP, podendo trazer benefícios importantes para a sociedade científica e tecnológica.

Comentários e Considerações sobre a Pesquisa:

O projeto tem mérito científico e, com certeza, a implementação desse novo sistema no HRAC facilitará o gerenciamento de dados o que, por sua vez, favorecerá o desenvolvimento de pesquisas futuras. A metodologia empregada no estudo utilizará, exclusivamente, dados secundários não havendo infrações éticas que impossibilitem a realização da pesquisa.

Endereço: SILVIO MARCHIONE 3-20
Bairro: VILA NOVA CIDADE UNIVERSITARIA **CEP:** 17.012-900
UF: SP **Município:** BAURU
Telefone: (14)3235-8421 **Fax:** (14)3234-7818 **E-mail:** uep_projeto@centrinho.usp.br



Continuação do Parecer: 1.753.467

Considerações sobre os Termos de apresentação obrigatória:

O pesquisador apresentou todos os termos obrigatórios: Carta de encaminhamento dos pesquisadores aos CEP; Formulário HRAC; Folha de Rosto Plataforma Brasil; Termo de Compromisso de Manuseio de Informações; Termo de Compromisso de Tornar Públicos os Resultados da Pesquisa e Destinação de Materiais ou Dados Coletados; Termo de Compromisso do Pesquisador Responsável. O pesquisador apresentou, ainda, ofício assinado pela Coordenadora Administrativa do Projeto Florida, confirmando estar ciente da realização da pesquisa nas dependências do ECR-PF no HRAC-USP e aprovando o acesso à base de dados do referido projeto para o desenvolvimento da pesquisa.

Recomendações:

Não há.

Conclusões ou Pendências e Lista de Inadequações:

Tendo em vista que o projeto não fere os princípios éticos sugiro ao CEP a sua aprovação.

Considerações Finais a critério do CEP:

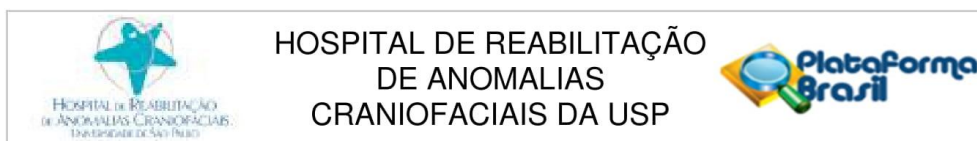
O pesquisador deve atentar que o projeto de pesquisa aprovado por este CEP refere-se ao protocolo submetido para avaliação. Portanto, conforme a Resolução CNS 466/12, o pesquisador é responsável por "desenvolver o projeto conforme delineado", se caso houver alterações nesse projeto, este CEP deverá ser comunicado em emenda via Plataforma Brasil, para nova avaliação.

Cabe ao pesquisador notificar via Plataforma Brasil o relatório final para avaliação. Os Termos de Consentimento Livre e Esclarecidos e/ou outros Termos obrigatórios assinados pelos participantes da pesquisa deverão ser entregues ao CEP. Os relatórios semestrais devem ser notificados quando solicitados no parecer.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Outros	71_2016_Checklist_Prot_Pesq.pdf	15/09/2016 13:59:39	Rafael Mattos de Deus	Aceito
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_735555.pdf	14/09/2016 17:22:34		Aceito
Projeto Detalhado / Brochura Investigador	DOCTORADO_Patrick_CEP_hrac_UPD_ATE.pdf	14/09/2016 17:20:41	Patrick Pedreira Silva	Aceito
Outros	TermodeCompromissoAtualizado.pdf	12/09/2016 11:34:00	Patrick Pedreira Silva	Aceito

Endereço: SILVIO MARCHIONE 3-20
Bairro: VILA NOVA CIDADE UNIVERSITARIA **CEP:** 17.012-900
UF: SP **Município:** BAURU
Telefone: (14)3235-8421 **Fax:** (14)3234-7818 **E-mail:** uep_projeto@centrinho.usp.br



Continuação do Parecer: 1.753.467

Folha de Rosto	Patrick_Folha_Rosto.pdf	09/09/2016 09:00:37	Patrick Pedreira Silva	Aceito
Outros	TornapublicoPatrick.pdf	18/08/2016 10:05:29	Patrick Pedreira Silva	Aceito
Outros	termouseoBanco.pdf	18/08/2016 10:01:48	Patrick Pedreira Silva	Aceito
Outros	ManuseioPatrick.pdf	18/08/2016 10:01:07	Patrick Pedreira Silva	Aceito
Outros	FormularioHRACPatrick.pdf	18/08/2016 10:00:49	Patrick Pedreira Silva	Aceito
Outros	CartaEncaminhamentoPatrick.pdf	18/08/2016 10:00:23	Patrick Pedreira Silva	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

BAURU, 29 de Setembro de 2016

Assinado por:
Silvia Maria Graziadei
(Coordenador)

Endereço: SILVIO MARCHIONE 3-20
Bairro: VILA NOVA CIDADE UNIVERSITARIA **CEP:** 17.012-900
UF: SP **Município:** BAURU
Telefone: (14)3235-8421 **Fax:** (14)3234-7818 **E-mail:** uep_projeto@centrinho.usp.br

Apêndices

**APÊNDICE A - DECLARAÇÃO DE USO EXCLUSIVO DE ARTIGOS EM
DISSERTAÇÃO/TESE**

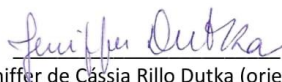
DECLARAÇÃO DE USO EXCLUSIVO DE ARTIGO EM DISSERTAÇÃO/TESE

Declaramos estarmos cientes de que o trabalho “Aplicação de mineração de dados para descoberta de conhecimento sobre fatores associados à ocorrência de fístulas após palatoplastia” será apresentado na Tese do aluno Patrick Pedreira Silva e que não foi e nem será utilizado em outra dissertação/tese dos Programas de Pós-Graduação do HRAC-USP.

Bauru, 10 de novembro de 2021.



Patrick Pedreira Silva (aluno)



Jeniffer de Cássia Rillo Dutka (orientadora)