UNIVERSIDADE FEEVALE

WESLLEI FELIPE HECKLER

ANÁLISE PREDITIVA SOBRE PACIENTES DO “PROJETO DE EXTENSÃO REABILITAÇÃO PULMONAR” DA UNIVERSIDADE FEEVALE

Novo Hamburgo

2018

WESLLEI FELIPE HECKLER

ANÁLISE PREDITIVA SOBRE PACIENTES DO “PROJETO DE EXTENSÃO REABILITAÇÃO PULMONAR” DA UNIVERSIDADE FEEVALE

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale.

Orientador: Dr. Juliano Varella de Carvalho

Novo Hamburgo

2018

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha família por sempre estarem ao meu lado me incentivando, me apoiando e me dando forças para realizar meus sonhos. Sem vocês nada disso seria possível.

Gostaria de agradecer ao meu orientador Juliano por toda a dedicação, auxílio e conhecimento proporcionado durante o desenvolvimento deste trabalho.

Também gostaria de agradecer à minha namorada Gabriela por toda a ajuda, incentivo e compreensão durante esse período.

Gostaria de agradecer aos meus amigos Feevaleiros pela parceria e irmandade de sempre.

Por fim, gostaria de agradecer a todos aqueles que de alguma forma contribuíram para a conquista deste objetivo.

A todos vocês, muito obrigado!

RESUMO

As doenças respiratórias atingem um dos principais sistemas do corpo humano e afetam grande parte da população brasileira. Nos casos mais graves, podem limitar as funcionalidades e a força muscular dos portadores, impactando em atividades cotidianas simples e, consequentemente, na qualidade de vida dos mesmos. Os programas de reabilitação pulmonar auxiliam no tratamento dessas doenças. O "Projeto de Extensão Reabilitação Pulmonar" da Universidade Feevale é um programa de reabilitação pulmonar que atende pacientes da comunidade e visa melhorar a qualidade de vida de portadores de doenças respiratórias crônicas através do desenvolvimento de ações educativas e assistenciais. As informações sobre os pacientes e resultados do tratamento são armazenadas em uma base de dados. A quantidade de dados dificulta a análise dos resultados. Neste trabalho, foi desenvolvida uma ferramenta para aplicar técnicas de *machine learning* na base de dados do "Projeto de Extensão Reabilitação Pulmonar". O objetivo da ferramenta é identificar tendências de abandono dos pacientes que estão ingressando no tratamento e extrair conhecimento sobre a base de dados para, com isso, contribuir na aplicação do tratamento de reabilitação pulmonar. Além disso, a ferramenta também disponibiliza visualizações para auxiliar na leitura dos dados e dos resultados por profissionais da área da saúde. Foram comparadas as técnicas *Support Vector Machine, Decision Tree* e *Random Forest*, onde *Random Forest* demonstrou melhor acurácia na predição de abandono.

Palavras-chave: *Machine learning.* Análise preditiva. Extração de conhecimento. Doenças respiratórias crônicas. Programas de Reabilitação Pulmonar.

ABSTRACT

Respiratory diseases reach one of the main systems of the human body and affect a large part of the Brazilian population. In severe cases, they can limit the functionalities and the muscular strength of the disease carriers, impacting on simple daily activities and, as result, their quality of life. Pulmonary rehabilitation programs help to treat these diseases. The "Pulmonary Rehabilitation Extension Project" of Feevale University is a pulmonary rehabilitation program that helps patients with chronic respiratory diseases in the community and aims to improve their quality of life through the development of educational and assistance actions. The information about the patients and the treatment results are stored in a database. The amount of data difficults the analysis of results. In this paper, it was developed a tool to apply machine learning techniques in the database of the "Pulmonary Rehabilitation Extension Project". The tool aims to identify the abandonment trends of patients entering the treatment and to extract knowledge about this database to contribute to the application of pulmonary rehabilitation treatment. In addition, the tool also provides visualizations to assist reading the data and results by healthcare professionals. The Support Vector Machine, Decision Tree and Random Forest techniques were compared, where the Random Forest technique demonstrated a better accuracy in the abandonment predictions.

Keywords: Machine learning. Predictive Analysis. Knowledge Discovery. Chronic Respiratory Diseases. Pulmonary Rehabilitation Programs.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AUC Área sob a curva ROC

DPOC Doença Pulmonar Obstrutiva Crônica

DRC Doença Respiratória Crônica

DT *Decision Tree*

GBT *Gradient Boosting Trees*

INCA Instituto Nacional de Câncer

KNN *K-Nearest Neighbour Classifier*

LR *Logistic Regression*

NB *Naïve Bayes*

PNCT Programa Nacional de Controle ao Tabagismo

PRP Programa de Reabilitação Pulmonar

RF *Random Forest*

ROC *Receiver Operating Characteristic*

SVM *Support Vector Machine*

TOS *Tree of Science*

SUMÁRIO

1 INTRODUÇÃO

Atualmente as doenças respiratórias atingem grande parte da população brasileira. Em 2015, o Ibope realizou uma pesquisa na qual revelou que 44% dos brasileiros apresentam sintomas de doenças respiratórias. A pesquisa também revelou que a maior incidência dos sintomas foi nos estados do sul do país (VIDALE, 2015).

Segundo Duchiade (1992), a poluição do ar é um dos principais fatores que causam as doenças respiratórias, podendo tanto atingir pessoas sadias quanto agravar os sintomas de pessoas que já possuem uma doença respiratória. Outro fator bastante significativo é o tabagismo, que é o principal agente causador da Doença Pulmonar Obstrutiva Crônica (DPOC) (SOCIEDADE BRASILEIRA DE PNEUMOLOGIA E TISIOLOGIA, 2004).

Essas doenças atingem um dos principais sistemas do corpo humano e, por isso, suas consequências são graves. Em níveis mais críticos, podem afetar atividades cotidianas simples. Segundo Poulain *et al.* (2003), pacientes com DPOC apresentam dificuldades na realização de exercícios, como por exemplo a caminhada, devido à obstrução das vias aéreas. Além disso, os pacientes portadores de DPOC também podem apresentar ansiedade e depressão, causando fadiga e dispneia. Esses sintomas implicam significativamente na realização de atividades diárias e, consequentemente, na qualidade de vida dos pacientes (MAURER *et al.*, 2008).

As doenças respiratórias são divididas em dois grupos conforme prejudicam as funções pulmonares. As doenças respiratórias restritivas têm como característica a redução da capacidade pulmonar (RODRIGUES *et al.*, 2002). Já as obstrutivas “são aquelas associadas a aumento da resistência nas vias aéreas, que se reflete funcionalmente por significativa redução nos fluxos expiratórios máximos” (FILHO, 1998). Também existe a doença respiratória mista, que é caracterizada pela junção dos sintomas das doenças restritivas e obstrutivas.

Um dos sintomas mais comuns das doenças respiratórias é a dispneia, uma sensação de falta de ar. Esse sintoma pode variar de acordo com cada caso, podendo ser mais intenso em alguns pacientes. Em geral, esse sintoma é percebido quando o paciente realiza alguma atividade, mesmo que essa atividade não exija muito esforço. Além disso, a ansiedade e o medo também podem causar essa sensação, visto que a falta de ar também está ligada ao emocional do paciente (GILMAN; BANZETT, 2009).

Os Programas de Reabilitação Pulmonar (PRPs) são bastante indicados para o tratamento e têm resultados muito positivos. Sobre esses programas, Nici *et al.* (2006, tradução nossa) comentam que:

> A reabilitação pulmonar é uma intervenção baseada em evidências, multidisciplinar e abrangente para pacientes com doenças respiratórias crônicas que são sintomáticas e muitas vezes diminuem as atividades da vida diária. Integrada no tratamento individualizado do paciente, a reabilitação pulmonar é projetada para reduzir os sintomas, otimizar o estado funcional, aumentar a participação e reduzir os custos dos cuidados de saúde através da estabilização ou reversão das manifestações sistêmicas da doença.

O "Projeto de Extensão Reabilitação Pulmonar" da Universidade Feevale é um PRP que atende pacientes da comunidade do Vale do Rio dos Sinos, de ambos os sexos e com idade superior a 40 anos. O projeto visa desenvolver e implementar ações educativas e assistenciais que promovam a melhoria da qualidade de vida em portadores de doenças respiratórias crônicas (DRC), tais como asma, DPOC, fibrose pulmonar, bronquiectasia, entre outras. O tratamento dura três meses em média.

São realizadas entrevistas pré e pós-tratamento por meio de formulários para registro de informações. As entrevistas são aplicadas para coletar informações sobre o perfil do paciente, como por exemplo gênero, idade, altura, peso, doença respiratória e período do tratamento. Além disso, visam coletar indicadores sobre a saúde, como por exemplo o número de vezes que o paciente tossiu em determinado período, a existência de secreção ou pressão no peito e a ocorrência de dispneia. Também são coletados indicadores sobre a qualidade de vida do paciente. Entre esses indicadores estão o grau de limitação de atividades habituais diárias (caminhar, conversar, arrumar a cama, lavar a louça, entre outras), a qualidade do sono e o nível de disposição.

Juntamente a essas informações, é registrado o desempenho dos pacientes em testes de carga máxima, que consistem na aplicação de 10 exercícios durante o tratamento de reabilitação. Os indicadores de desempenho são utilizados para medir a evolução do paciente durante o tratamento. Essas informações são registradas desde 2002 em uma base de dados armazenada em uma planilha Excel. Ao todo, a base de dados contém 542 linhas, que representam os pacientes, e 356 colunas, que representam os atributos. A quantidade de registros e atributos dificulta a análise e a visualização dos resultados. Em contrapartida, é possível que a base de dados contenha informações importantes sobre relacionamentos entre pacientes e doenças respiratórias que podem ser estudadas por profissionais da saúde.

A quantidade de dados no mundo está aumentando e tende a continuar crescendo. Quanto maior essa quantidade, mais difícil se torna o entendimento das pessoas sobre esses dados, que podem conter informações potenciais. Desde que a vida humana começou, as pessoas buscam padrões sobre os dados, visando obter algum benefício. Caçadores buscam padrões sobre a migração de animais, agricultores buscam padrões sobre o crescimento das plantações e empreendedores buscam padrões de comportamento que podem ser transformados em negócios lucrativos. Padrões podem trazer explicações sobre os dados e auxiliar na predição de dados futuros (WITTEN; FRANK, 2005).

Extrair conhecimento sobre bases de dados pode gerar informações úteis e benefícios para quem as analisa. Goldschmidt e Passos (2005) destacam que:

> A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Portanto, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto de aplicação.

Segundo Mitchell (2006), *Machine Learning* abrange as áreas de Ciência da Computação e Estatística e consiste na criação de algoritmos que aprendam e se adaptem automaticamente a partir da experiência adquirida sobre um determinado problema. Ademais, também consiste no estudo das inferências que podem ser realizadas sobre os dados analisados.

Técnicas de *Machine Learning* podem ser aplicadas na área da saúde de diversas formas. Alguns exemplos de aplicação são a classificação de imagens microscópicas de células, detecção de surtos de doenças e detecção de padrões anômalos de sintomas e sua distribuição geográfica (MITCHELL, 2006).

A análise preditiva é uma subárea de *Machine Learning* e tem como objetivo prever comportamentos futuros a partir do aprendizado sobre dados do passado referentes a determinado problema. Conforme Alpaydin (2010, tradução nossa) destaca, “assim que tivermos uma regra que se ajusta aos dados anteriores, se o futuro for semelhante ao passado, podemos fazer previsões corretas para novas instâncias”.

Essa técnica tem diversas aplicações na área da saúde. Como, por exemplo, na análise de quais tratamentos serão mais eficientes para determinados futuros pacientes (MITCHELL, 2006) e na busca por doadores de sangue (SILVA, 2018). Outro exemplo de aplicação é na predição da ação de medicamentos através da análise de suas propriedades químicas e de sua

estrutura tridimensional, acelerando a descoberta de novos medicamentos e reduzindo seu custo (WITTEN; FRANK, 2005).

Este trabalho, portanto, propõe o desenvolvimento de uma ferramenta para extração de conhecimento da base de dados do “Projeto de Extensão Reabilitação Pulmonar”. Por meio da aplicação de *Machine Learning*, a ferramenta tem como objetivo identificar a tendência de abandono dos pacientes que estão ingressando no tratamento. As técnicas *Support Vector Machine*, *Decision Tree* e *Random Forest* foram comparadas, a fim de identificar a melhor técnica para a predição de abandono.

A ferramenta também visa facilitar a leitura dos resultados através da geração de visualizações sobre os dados. Ela é formada por um conjunto de painéis de visualização interativos, onde os resultados da análise preditiva são apresentados em forma de tabela. Nos painéis, também são gerados gráficos que exibem estatísticas sobre a base de dados, tais como quantidade e percentual geral de abandono, proporção de doenças, distribuição de idade por gênero, proporção de escala de dispneia, média de internações por doença, quantidade de pacientes por comorbidade, quantidade de abandonos por gênero e quantidade de abandonos por escala de dispneia. Também são geradas visualizações para identificação de tendências referentes aos perfis dos pacientes da base de dados, como por exemplo dispersão de espirometria VEF1 % por idade e escala de dispneia por espirometria VEF1 % e gênero.

Este trabalho está dividido em 7 capítulos. O primeiro capítulo apresentou uma introdução sobre o trabalho. O segundo capítulo apresenta conceitos sobre reabilitação pulmonar e o “Projeto de Extensão Reabilitação Pulmonar”. Uma revisão sistemática sobre técnicas de análise preditiva aplicadas à reabilitação pulmonar é apresentada no terceiro capítulo. O quarto capítulo aborda a análise preditiva, expondo algumas de suas técnicas. Os resultados da comparação das técnicas escolhidas são abordados no quinto capítulo. O sexto capítulo descreve o desenvolvimento da ferramenta. Por fim, o último capítulo apresenta as conclusões deste trabalho.

2 REABILITAÇÃO PULMONAR

A reabilitação pulmonar é um componente de caráter preventivo e terapêutico utilizado no tratamento de pacientes com DRC. Essa é uma estratégia de tratamento complementar e não substitui outros tratamentos médicos. A reabilitação pulmonar é um programa individualizado e o tratamento deve ser planejado para atender às necessidades individuais de cada paciente (AMERICAN ASSOCIATION OF CARDIOVASCULAR AND PULMONARY REHABILITATION, 2007).

Os principais objetivos dos PRPs são diminuição de sintomas, melhoria na qualidade de vida, aumento da tolerância ao exercício, maior independência nas atividades de vida diária e diminuição do uso de medicamentos. Também existem objetivos específicos que são estabelecidos pelo próprio paciente em conjunto com os profissionais da saúde. Esses últimos são utilizados como forma de motivação do paciente durante a execução do tratamento. Em geral, os objetivos do paciente são melhorar a respiração, ser mais ativo, diminuir a ansiedade e a depressão, ser mais independente e autoconfiante, melhorar a qualidade de vida e ser capaz de realizar normalmente atividades cotidianas simples, tais como limpar a casa, tomar banho e executar passatempos. Desta forma, os PRPs visam atingir os objetivos do programa e do paciente durante o tratamento (AMERICAN ASSOCIATION OF CARDIOVASCULAR AND PULMONARY REHABILITATION, 2007).

Está comprovado que os PRPs podem gerar resultados bastante significativos para os pacientes participantes. Entre os resultados estão a redução de dispneia e fadiga, aumento da capacidade de realizar exercícios e atividades diárias, melhora na qualidade de vida relacionada à saúde, redução de ansiedade e depressão, redução da hospitalização e do uso de medicamentos e o retorno de alguns pacientes ao trabalho. Sendo assim, a reabilitação pulmonar é bastante indicada para pessoas com doenças respiratórias crônicas (AMERICAN ASSOCIATION OF CARDIOVASCULAR AND PULMONARY REHABILITATION, 2007).

2.1 DOENÇAS RESPIRATÓRIAS CRÔNICAS

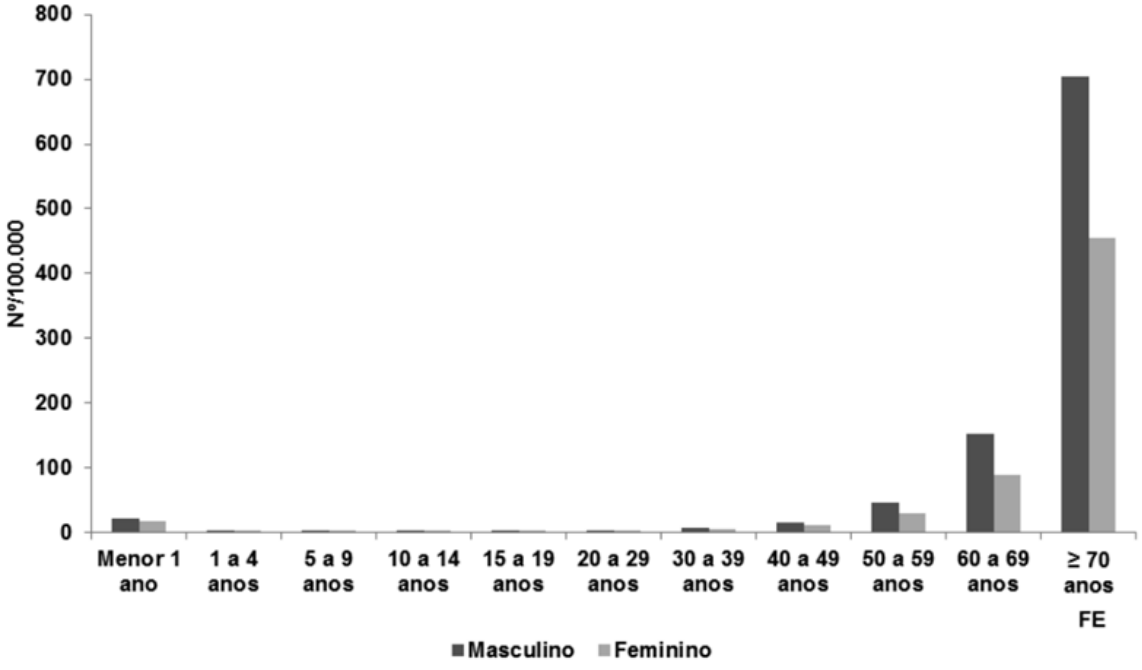
As DRCs são divididas em dois grupos de acordo com o impacto que causam nas funções pulmonares de seus portadores. Elas podem ser classificadas como **restritivas** ou **obstrutivas**. Existem também doenças que unem características de ambos os grupos

simultaneamente, sendo estas classificadas por doenças respiratórias **mistas** (RODRIGUES *et al.*, 2002).

As DRCs representam cerca de 7% da mortalidade global, causando 4,2 milhões de óbitos anuais. A DPOC afeta mais de 200 milhões de pessoas em todo o mundo (GOULART, 2011). No Brasil, essas doenças foram classificadas como a terceira maior causa de morte relacionada à doenças crônicas não transmissíveis em 2011 (MALTA *et al.*, 2014).

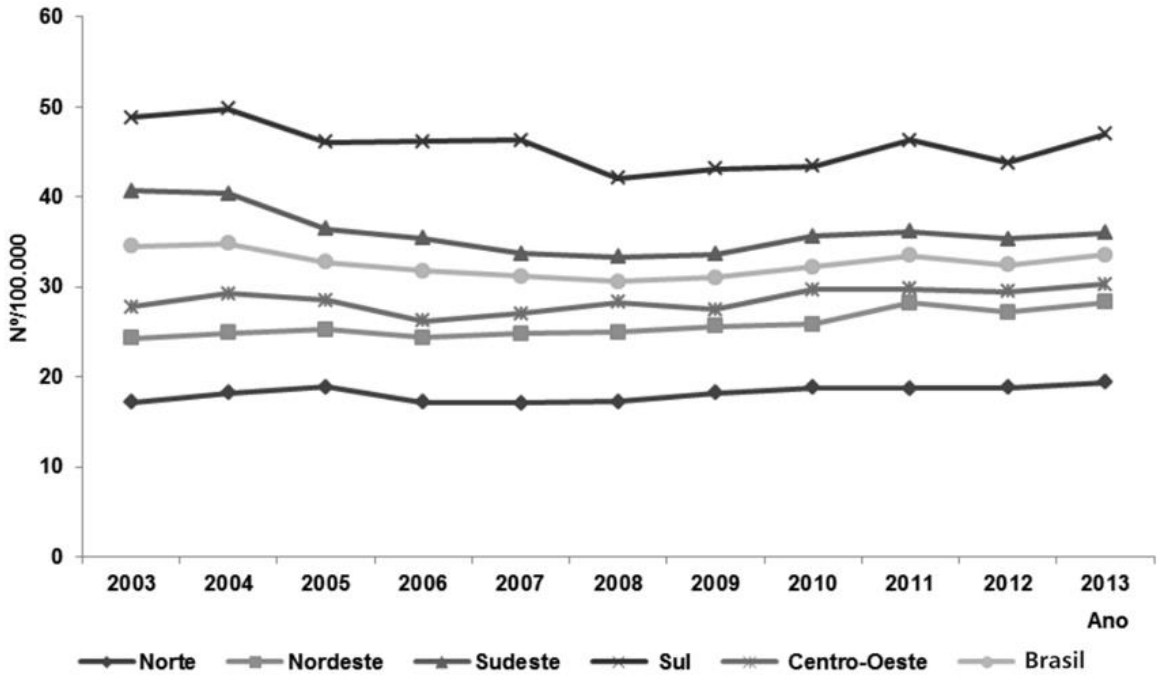
Entre 2003 e 2013, foram registrados 685.031 óbitos por DRC no Brasil. As principais causas de óbitos relacionados à DRC foram a bronquite, o enfisema e DPOC (SECRETARIA DE VIGILÂNCIA EM SAÚDE, 2016). A Figura 1 apresenta a média das taxas de mortalidade por DRC no Brasil conforme sexo e faixa etária entre os anos de 2003 e 2013. Nesta figura, pode-se observar que a incidência de óbitos é maior em indivíduos com idade igual ou superior a 70 anos. A Figura 2 apresenta as taxas de mortalidade por DRC separadas por regiões no Brasil no período de 2003 à 2013. Pode-se observar que a maior taxa de mortalidade é referente aos estados do sul do país.

Figura 1 - Média das taxas de mortalidade por doença respiratória crônica, segundo sexo e faixa etária, Brasil, 2003 a 2013.



Fonte: Secretaria de Vigilância em Saúde (2016)

Figura 2 - Taxas de mortalidade por doença respiratória crônica por regiões, Brasil, 2003 a 2013.



Fonte: Secretaria de Vigilância em Saúde (2016)

2.1.1 Doenças restritivas

As doenças restritivas são caracterizadas pela redução dos volumes pulmonares. A gravidade da doença tende a aumentar com o avanço da idade dos portadores (SPERANDIO *et al.*, 2016). Barreto (2002) complementa que essas doenças podem ser classificadas pela "redução da capacidade pulmonar total por anormalidades hipodinâmicas, neurais ou musculares". A fibrose pulmonar é um exemplo de pneumopatia com padrão restritivo (COSTA; JAMAMI, 2001).

O diagnóstico é realizado por meio da medição dos volumes pulmonares estáticos (SPERANDIO *et al.*, 2016). No diagnóstico, é sugerida a existência de distúrbio restritivo sempre que houver redução da capacidade vital não explicada por distúrbio obstrutivo. A capacidade vital compreende o volume de ar corrente, que é o volume de ar inspirado e expirado espontaneamente em cada ciclo respiratório, e os volumes inspiratório e expiratório de reserva, que são, respectivamente, os volumes máximos que podem ser inspirados ou expirados voluntariamente ao final de uma inspiração ou expiração espontânea. A redução da

capacidade pulmonar total pode ser a única alteração relacionada à esse grupo de doenças (BARRETO, 2002).

2.1.2 Doenças Obstrutivas

As doenças obstrutivas são caracterizadas pela redução do fluxo de ar máximo que pode ser deslocado do pulmão. Essa redução ocorre devido ao estreitamento das vias aéreas durante a exalação (PELLEGRINO *et al.*, 2005). Um quadro de obstrução pulmonar pode não ser reversível dependendo de sua gravidade. O enfisema pulmonar e a bronquite crônica são exemplos de doenças que possuem padrão obstrutivo (COSTA; JAMAMI, 2001).

Uma das principais formas de diagnóstico da obstrução ao fluxo aéreo é a medida do volume expiratório forçado no primeiro segundo (SOCIEDADE PAULISTA DE PNEUMOLOGIA E TISIOLOGIA, 2008), que representa o volume máximo que um indivíduo consegue expirar no primeiro segundo de uma expiração máxima. Esse parâmetro é utilizado para medir o fluxo aéreo da maior parte das vias aéreas (COSTA; JAMAMI, 2001). Os indicadores de exacerbações, dispneia, mortalidade, hiperinsuflação, qualidade de vida e capacidade de exercícios também são utilizados para o diagnóstico (SOCIEDADE PAULISTA DE PNEUMOLOGIA E TISIOLOGIA, 2008).

2.1.3 Causas/Fatores de Risco

Indivíduos fumantes são mais propensos a contrair doenças respiratórias crônicas. O fumo possui aproximadamente 7000 substâncias químicas e é o maior tóxico que o ser humano introduz no organismo. O tabagismo é a mais importante causa de DPOC e é uma das principais causas de outras DRCs (TARANTINO, 2002). Além do tabagismo, os alergênicos, os agentes ocupacionais, os fatores genéticos, sociais e relacionados ao estilo de vida também são fatores de risco para DRC. As DRCs tendem a se agravar conforme o envelhecimento do paciente (SECRETARIA DE VIGILÂNCIA EM SAÚDE, 2016).

A união de variações acentuadas na temperatura e baixa umidade, que ocorre em algumas regiões do Brasil, em determinadas épocas do ano, pode causar sintomas respiratórios. O ar seco influencia a função mucociliar, que é o principal mecanismo de defesa do sistema respiratório e é responsável por filtrar poeira, bactérias e vírus inspirados do ar. Durante o dia, um indivíduo inspira cerca de 10.000 a 20.000 litros de ar. Isso, consequentemente, torna o sistema respiratório fácil de ser infiltrado. Portanto, quando esse

mecanismo é comprometido, torna maior a possibilidade de infecção respiratória (SOCIEDADE PAULISTA DE PNEMOLOGIA E TISIOLOGIA, 2008).

A poluição atmosférica também causa impactos sobre o sistema respiratório. O aumento no número de veículos automotores e o crescimento de áreas industrializadas geram uma grande concentração de poluentes no ar e, por consequência, podem gerar problemas graves em relação à qualidade do ar. Além de gerar custos diretos com tratamentos de doenças respiratórias, a poluição também pode gerar custos indiretos com diminuição da assiduidade no trabalho e queda na produtividade. Dentre os fatores de risco, a poluição é um dos poucos que é passível de mudança, já que pode ser amenizada por políticas de controle e prevenção (SOCIEDADE PAULISTA DE PNEUMOLOGIA E TISIOLOGIA, 2008).

2.1.4 Sintomas

Os principais sintomas relatados por pacientes com DPOC são fadiga e dispneia. Esses sintomas ocorrem devido ao aumento do consumo de oxigênio, da ventilação pulmonar e produção de dióxido de carbono. Conforme a progressão da doença, os sintomas se tornam mais recorrentes, impactando em atividades de vida diária como pentear os cabelos, trocar de roupa ou ao cuidar da higiene pessoal (REGUEIRO *et al.*, 2006).

A dispneia é um dos sintomas apresentados por doenças respiratórias e consiste na dificuldade de respirar. Ela pode ocorrer durante a prática de um exercício que exige grande esforço, durante a prática de alguma atividade de rotina e até mesmo durante o repouso (TARANTINO, 2002). A dispneia é um sintoma que pode ser ocasionado por diversos fatores, como o lado emocional do paciente por exemplo. O sentimento de fragilidade ou de medo pode causar sensação de dispneia nos pacientes (SOCIEDADE PAULISTA DE PNEUMOLOGIA E TISIOLOGIA, 2008). Além de dispneia, pacientes com doenças respiratórias também apresentam dor torácica, tosse, ruídos na expiração chamados de sibilância, rouquidão ou alteração no timbre da voz e cornagem, que consiste na dificuldade inspiratória por redução do calibre das vias respiratórias superiores (TARANTINO, 2002).

2.1.5 Consequências

Os indivíduos com doença respiratória crônica possuem limitações para a execução de atividades diárias devido à intolerância ao exercício causada pela doença. Os principais sintomas que causam essas limitações são dispneia e fadiga (NICI *et al.*, 2006).

Bauerle, Chrusch e Younes (1998) acrescentam que a tolerância ao exercício pode variar entre pacientes com DPOC. A incapacidade física de cada paciente está relacionada à sua função muscular inspiratória. Com um baixo fluxo inspiratório máximo, os músculos respiratórios ficam enfraquecidos e pode ser gerada uma alta resistência respiratória, colocando os músculos respiratórios sob estresse quando exigidos. Um determinado exercício pode exigir uma quantidade maior de ventilação, demandando, consequentemente, um esforço maior dos músculos inspiratórios, que muitas vezes é excessivo em relação à capacidade desses músculos.

Segundo um estudo realizado por Sperandio *et al.* (2016), o distúrbio ventilatório restritivo está relacionado com a falta de equilíbrio independentemente de idade, gênero ou comorbidades. Desta forma, o risco de quedas é maior em pacientes com doença restritiva. Em consequência disso, pacientes com esse tipo de doença tendem a ter limitações nas atividades de vida diária.

A alteração na função pulmonar e a dispneia, sintomas da DPOC, podem levar as consequências das DRC a um nível psicológico. A intolerância ao exercício ocasionada por esses sintomas gera limitações até mesmo para atividades cotidianas diárias. Essa limitação pode despertar ansiedade e depressão em pacientes ao se verem incapazes de realizar tarefas simples do cotidiano (COSTA *et al.*, 2014). Jennings *et al. (*2009) destacam uma associação de sintomas de depressão com diminuição da qualidade de vida, maior tempo de internação hospitalar, aumento da carga de sintomas, maior readmissão hospitalar e até mortalidade em pacientes portadores de DPOC.

2.2 PROGRAMAS DE REABILITAÇÃO PULMONAR

A reabilitação pulmonar, segundo Palombini *et al.* (2001), "é uma abordagem multidisciplinar que tem como objetivo melhorar a qualidade de vida do paciente portador de doença pulmonar crônica". Nici *et al.* (2006, tradução nossa) acrescentam que

> Os programas de reabilitação pulmonar envolvem avaliação do paciente, treinamento físico, educação, intervenção nutricional e apoio psicossocial. Em um sentido mais amplo, a reabilitação pulmonar inclui um espectro de estratégias de intervenção integradas ao manejo vitalício de pacientes com doença respiratória crônica e envolve uma colaboração dinâmica e ativa entre o paciente, a família e os profissionais de saúde.

A equipe responsável pelo programa deve ser formada por um pneumologista, fisioterapeutas, enfermeiras, psicólogos e/ou psiquiatras, nutricionistas, terapeutas

ocupacionais e assistentes sociais. O fisioterapeuta é responsável pelo monitoramento do exercício, treinamento respiratório e treinamento sobre conservação de energia. O nutricionista é encarregado da avaliação e orientação dietética. O assistente social orienta e adapta as atividades de vida diária. Psicólogo e psiquiatra são responsáveis pelo tratamento da depressão ou outras alterações. A equipe também deve contar com um coordenador com experiência no atendimento de pacientes e um médico para organizar o programa. Com isso, os programas de reabilitação abrangem todos os aspectos necessários para o tratamento do paciente, desde as medicações até as ações educativas a fim de incentivar a mudança comportamental do paciente e prevenir as doenças respiratórias (PALOMBINI *et al.*, 2001).

Além da multidisciplinaridade, os programas também devem ser adaptados às necessidades de cada paciente. A limitação física pode variar para cada paciente conforme a gravidade da DRC. Diante disso, os pacientes devem receber avaliação individual e os programas devem ser planejados para atingir metas individuais e realistas (RIES *et al.*, 2007). O tratamento visa permitir que os pacientes sejam independentes, permitindo com que consigam cuidar de si mesmos, exercer as atividades diárias e que sejam menos dependentes de profissionais da saúde. Portanto, tem como objetivo básico melhorar a qualidade de vida dos pacientes (TARANTINO, 2002).

Embora a reabilitação seja tradicionalmente direcionada para pacientes com DPOC, ela é indicada para todos os pacientes com pneumopatias que continuam apresentando dispneia mesmo recebendo tratamento médico adequado (TARANTINO, 2002). Ela não deve ser aplicada somente em pacientes com estado mais grave de doença respiratória crônica, pois pode ser benéfica para todos os pacientes que possuem capacidade pulmonar reduzida, ou cuja qualidade de vida é afetada de alguma forma devido aos sintomas das doenças. Desta forma, deve fazer parte do tratamento de qualquer paciente que possua alguma doença respiratória crônica, abordando seus déficits funcionais e/ou psicológicos (NICI *et al.*, 2006).

Apesar de o tabagismo ser uma das principais causas de doenças respiratórias crônicas, ainda é debatida a indicação do tratamento de reabilitação para pacientes fumantes. Entende-se que o ato de fumar representa uma incapacidade de autoajuda do paciente e uma influência para o abandono do tratamento. É necessário que o paciente esteja motivado, esperançoso e tenha conhecimento sobre os benefícios do tratamento e suas limitações físicas. Do contrário, o tratamento pode não gerar os resultados desejados (TARANTINO, 2002).

Muitos pacientes ingressam em PRP com um baixo condicionamento físico, um fator bastante relevante na piora da qualidade de vida. Desta forma, um dos principais aspectos da reabilitação pulmonar é a melhora no condicionamento físico para, assim, melhorar a condição cardiovascular. Esse processo passa pela aplicação de exercícios dirigidos às extremidades superiores, inferiores e ao sistema cardiovascular. Nos exercícios aplicados para melhora das extremidades inferiores e da condição cardiovascular, é indicado o uso de esteira ou bicicleta ergométrica com acompanhamento técnico. Já para as extremidades superiores, é indicado trabalhar a musculatura acessória da respiração, como por exemplo o diafragma (PALOMBINI *et al.*, 2001). Segundo a American Association of Cardiovascular and Pulmonary Rehabilitation (2007), "é mais benéfico direcionar o treinamento com exercícios para os músculos envolvidos nas atividades de vida diária".

Antes da aplicação do treinamento físico, deve ser realizada uma avaliação para analisar a tolerância a exercícios do paciente (AMERICAN ASSOCIATION OF CARDIOVASCULAR AND PULMONARY REHABILITATION, 2007). O teste de caminhada de seis minutos é utilizado para medir a capacidade física do paciente. Esse teste é um dos mais utilizados para essa avaliação por ser simples, reprodutível e possuir baixo custo. Devido a sua simplicidade, o teste pode ser realizado por qualquer profissional da saúde (TARANTINO, 2002).

O teste consiste em uma caminhada em linha reta durante um período de seis minutos, onde o paciente deve percorrer a maior distância possível. O teste deve sempre ser realizado em uma superfície plana como um corredor ou uma quadra de esportes. Dois examinadores devem acompanhar o paciente e, durante a execução do teste, medir alguns indicadores de saúde do paciente como, por exemplo, a frequência cardíaca, frequência respiratória, pressão arterial, índice de dispneia de Borg, entre outros. Estudos comprovam que o estímulo verbal dos examinadores pode incentivar o paciente, fazendo com que ele atinja uma distância maior. O teste deve ser aplicado pré e pós tratamento e o indicador de desempenho é a distância percorrida pelo paciente durante o teste. Uma maior distância percorrida na execução do teste após o tratamento indica uma melhora no estado de saúde do paciente (TARANTINO, 2002).

Em relação à qualidade de vida relacionada à saúde, são aplicados questionários para medir os resultados nessa área. Os questionários são respondidos pelos próprios pacientes e, em geral, apresentam múltiplas dimensões, avaliando estado funcional, sintomas (em especial dispneia e fadiga), domínio sobre a doença, impacto da doença sobre a pessoa e satisfação

geral ou insatisfação com a vida. Os questionários devem ser curtos e de fácil leitura pelo paciente. Além disso, devem ter boas propriedades avaliativas e de fácil administração pela equipe do programa. *St. George's Respiratory Questionnaire* e *Ferrans & Powers QOL – Pulmonary Version* são exemplos de questionários aprovados sobre qualidade de vida respiratória e podem ser úteis para PRP (AMERICAN ASSOCIATION OF CARDIOVASCULAR AND PULMONARY REHABILITATION, 2007).

A avaliação dos resultados do programa é um componente de grande importância para medir a eficiência do tratamento de cada paciente e do programa como um todo. A avaliação deve ocorrer pré e pós-reabilitação e deve incluir pelo menos as medidas de dispneia e tolerância ao exercício e as condições clínica e funcional (PALOMBINI *et al.*, 2001). A American Association of Cardiovascular and Pulmonary Rehabilitation (2007) acrescenta que esse processo não deve ser intimidante, representar sobrecarga ou consumir tempo excessivo.

Um estudo realizado por Zanchet, Viegas e Lima (2005) comprovou que a reabilitação pulmonar melhorou a distância percorrida no teste de caminhada de seis minutos, bem como a qualidade de vida e a capacidade de exercício funcional de pacientes com DPOC. Segundo Ries *et al.* (2007), a reabilitação pulmonar melhora a tolerância ao exercício, o estado de saúde e a qualidade de vida de pacientes com diversas DRCs, tais como DPOC, asma, bronquiectasia, fibrose cística, entre outras. Outro estudo proposto por Costa *et al.* (2014) constatou que houve uma melhora na qualidade de vida de pacientes portadores de DPOC com a diminuição da ansiedade e depressão após participação em um PRP.

2.3 "PROJETO DE EXTENSÃO REABILITAÇÃO PULMONAR" DA UNIVERSIDADE FEEVALE

O "Projeto de Extensão Reabilitação Pulmonar" da Universidade Feevale é um PRP criado em 2002 que propõe o desenvolvimento de ações na perspectiva da promoção da saúde e da melhoria da qualidade de vida. Para isso, o projeto propõe desenvolver ações educativas de promoção da saúde na área de prevenção e reabilitação de pacientes com DRC em parceria com um centro de especialidades da Prefeitura de Novo Hamburgo, município situado no estado do Rio Grande do Sul. São desenvolvidas intervenções terapêuticas que promovem uma melhoria na qualidade de vida dos indivíduos participantes. São realizadas ações interdisciplinares dos cursos de Educação Física, Fisioterapia, Nutrição, Medicina e

Psicologia. O projeto visa reinserir pessoas portadoras de DRC no convívio social a partir da melhora na qualidade de vida de seus beneficiados após a participação no projeto.

O programa é direcionado para pessoas de ambos os sexos, com idade superior a 40 anos de idade. Os pacientes devem estar obrigatoriamente acompanhados pelo Programa Nacional de Controle ao Tabagismo (PNCT) da Casa da Vacina de Novo Hamburgo ou apresentar diagnóstico de DRC. O PNCT é um programa do Ministério da Saúde realizado através do Instituto Nacional de Câncer (INCA) e tem como objetivo reduzir a prevalência de fumantes e a consequente morbimortalidade relacionada ao consumo de derivados de tabaco. A triagem de pacientes pelo PNCT ocorre para incentivar os pacientes a abandonar o tabaco.

Para mensuração dos resultados, o projeto possui objetivos definidos. O principal objetivo é o desenvolvimento de ações para possibilitar uma melhora na qualidade de vida a indivíduos tabagistas ou portadores de DRC através da melhora no condicionamento físico e do conhecimento de aspectos que envolvem a doença e seu tratamento. Além do objetivo geral, o projeto possui objetivos específicos que envolvem, além da melhora do condicionamento físico, otimizar o estado nutricional do paciente, reduzir a sensação de dispneia, melhorar a funcionalidade, reduzir exacerbações e internações hospitalares e reduzir o índice de tabagismo.

O projeto prevê o atendimento de 60 pacientes por ano, sendo 15 concomitantemente. Devido a infraestrutura do programa, essa é a capacidade máxima de atendimento. O encaminhamento de pacientes é realizado pela rede pública ou privada da região do Vale dos Sinos. Antes de ingressar no projeto, os pacientes são avaliados por um médico para confirmação de diagnóstico e suas condições clínicas. Após essa avaliação, o projeto é dividido em três fases. A primeira fase é a de avaliações. Nessa fase, são realizadas as avaliações médica, cardiopulmonar, psicológica, de dispneia, de impactos nas atividades de vida diária, física, nutricional e antropométrica. A segunda fase é a do treinamento físico. Nessa fase, são aplicados os exercícios para treinamento físico dos pacientes. Todos os pacientes passam por uma semana de adaptação ao treinamento físico, na qual realizam exercícios na academia em equipamentos sem carga. Após a adaptação, os pacientes realizam os exercícios com 50%, 60%, 70% e 80% da carga máxima, a qual é definida individualmente conforme as limitações de cada paciente. Nessa fase, também são aplicadas palestras para disseminar o conhecimento sobre DRC e realizados grupos de apoio da psicologia. Na terceira

fase, são realizadas as reavaliações dos pacientes, repetindo os procedimentos realizados na primeira fase.

O tratamento tem duração de 12 semanas com 3 atendimentos por semana. Nos atendimentos são realizadas atividades de condicionamento físico e palestras educacionais para pacientes, familiares e grupo de apoio. Todas as atividades envolvidas no projeto ocorrem nos campi da Universidade Feevale. Antes do início de cada atendimento diário, são medidos os sinais vitais do paciente, tais como frequência cardíaca, frequência respiratória, saturação periférica de oxigênio e pressão arterial. Em seguida é iniciado o treinamento de condicionamento físico. As palestras educativas e o grupo de apoio da psicologia ocorrem uma vez por semana.

Alguns formulários são utilizados para medir indicadores de saúde, qualidade de vida e o desempenho dos pacientes nos exercícios de condicionamento físico. Conforme o preenchimento dos formulários, os dados são registrados em uma base de dados armazenada em uma planilha Excel. Ao todo, a base de dados contém 542 linhas, que representam os pacientes, e 356 colunas, que representam os atributos. Os seguintes formulários são utilizados:

- *COPD Assessment Test*: Indicador de bem-estar do paciente. Este questionário é exclusivo para pacientes portadores de DPOC;
- Escala de Dispneia (*Medical Research Council*): Indicador de dispneia;
- Escala de Lócus de Controle da Saúde: Indicador sobre conhecimento do paciente sobre sua saúde;
- Índice Basal de Dispneia (IBD): Indicador de comprometimento funcional;
- Teste de Caminhada de 6 Minutos: Indicador de desempenho no teste de caminhada de 6 minutos;
- Ficha *London*: Indicador de falta de ar na execução de atividades de vida diária;
- Índice Transacional de Dispneia (ITD): Indicador de mudança do comprometimento funcional;
- *St. George's Respiratory Questionnaire*: Indicador de dificuldade respiratória em atividades de rotina;

- Teste de Carga Máxima: Indicador de desempenho nos exercícios aplicados no treinamento físico.

Devido à abordagem multidisciplinar dos PRPs, o projeto possui diversas equipes envolvidas, cada uma com suas respectivas atividades. A equipe da Fisioterapia é responsável pela realização do teste da caminhada dos seis minutos, da oximetria digital de pulso e pela espirometria. A equipe da Nutrição é responsável pela avaliação nutricional a partir de um recordatório alimentar e uma avaliação com a utilização de bioimpedância, além de acompanhamento das dietas. A equipe da Psicologia aplica o questionário de qualidade de vida (Questionário respiratório de *St. George's*, através de uma entrevista individual), sendo que este instrumento é específico para portadores de DRC, validado no Brasil. Além de realizarem os grupos de apoio, aplicam o inventário BECK (questionário de ansiedade/depressão) e realizam atendimentos individuais conforme as necessidades dos pacientes. A equipe da Educação Física realiza o teste de exercício cardiopulmonar, o teste de carga máxima e o acompanhamento dos exercícios na academia. O teste do exercício cardiopulmonar compreende a prescrição do treinamento aeróbico e muscular. Todas estas avaliações serão realizadas no início e no término do projeto.

O projeto prevê atingir algumas metas quantitativas. As metas têm correlação direta com os objetivos a serem alcançados. Dentre as metas definidas, estão a melhora em 20% do condicionamento físico, redução do consumo de tabaco em 20%, diminuição de exacerbações e internações hospitalares em 20%, a melhora na sensação de dispneia em 25%, otimização em 10% da funcionalidade dos pacientes para a realização de suas atividades de vida diária, melhora de 5% no estado nutricional e melhora de 4% na qualidade de vida.

2.4 CONSIDERAÇÕES FINAIS

As DRCs representam uma das principais causas de morte no mundo. No Brasil, foram registrados 685.031 óbitos pela doença entre 2003 e 2013. Essas doenças são divididas em dois grupos de acordo com sua manifestação no sistema respiratório. As doenças restritivas são caracterizadas pela redução dos volumes pulmonares. As doenças obstrutivas são caracterizadas pela obstrução das vias aéreas, reduzindo o fluxo de ar. Também existem doenças que possuem ambos os padrões, sendo essas denominadas doenças mistas. São vários os fatores de risco e causas dessas doenças, sendo o tabagismo o principal fator de risco. Essas doenças também geram grandes consequências na vida de seus portadores, desde

sintomas, tais como dispneia e tosse, até limitações de funcionalidades do paciente, diminuindo a tolerância ao exercício e, consequentemente, reduzindo a qualidade de vida.

Os PRPs são programas multidisciplinares indicados como tratamento complementar para DRC. Eles têm como principais objetivos a redução de sintomas e a melhora na qualidade de vida dos pacientes por meio de ações educativas e psicológicas e treinamentos físicos. É comprovado que os PRPs geram excelentes resultados no tratamento de pacientes com DRC. O “Projeto de Extensão Reabilitação Pulmonar” da Universidade Feevale é um PRP que propõe o desenvolvimento de ações que visam a promoção da saúde e da melhoria da qualidade de vida dos pacientes participantes. O projeto atende pacientes da comunidade do município de Novo Hamburgo encaminhados pela rede pública ou privada da região do Vale dos Sinos.

3 REVISÃO SISTEMÁTICA SOBRE TÉCNICAS DE ANÁLISE PREDITIVA APLICADAS À REABILITAÇÃO PULMONAR

Há fortes indícios de que o material relacionado à aplicação de técnicas de análise preditiva na área de reabilitação pulmonar é escasso, dificultando o desenvolvimento de pesquisas na área. Foi realizada uma revisão sistemática, visando verificar na literatura as principais técnicas de análise preditiva aplicadas a programas de reabilitação pulmonar. Além disso, também foram identificados os algoritmos, formas de validação, linguagens de programação e ferramentas/*frameworks* utilizados com maior frequência nessa área.

Com base nos resultados validados ao final dessa revisão, foram selecionadas técnicas, linguagem de programação e ferramentas/*frameworks* para a utilização neste trabalho. Foi realizado um estudo sobre cada escolha, visando aprofundar os conhecimentos sobre elas. Esse estudo é apresentado no Capítulo 4.

3.1 PROTOCOLO DA REVISÃO SISTEMÁTICA

Esta revisão sistemática foi desenvolvida com base no protocolo proposto por (MEDEIROS, 2016). Esse protocolo foi elaborado conforme o protocolo da pesquisadora Elisabete Kitchenham (KITCHENHAM, 2007) e o protocolo de recomendação PRISMA (PRISMA, 2018), que atendem às áreas da computação e saúde, respectivamente.

3.1.1 O Protocolo

O planejamento desta revisão sistemática foi realizado seguindo a estrutura:

a) Título

Revisão sistemática sobre técnicas de análise preditiva aplicadas à reabilitação pulmonar

b) Resumo

O protocolo desta revisão sistemática foi desenvolvido com base no protocolo elaborado por (MEDEIROS, 2016). O principal objetivo de estudo dessa revisão sistemática é buscar na literatura técnicas de análise preditiva aplicadas a programas de reabilitação pulmonar. O protocolo foi desenvolvido com ajuda de especialistas das áreas de computação e saúde.

c) Objetivo

Nesta revisão, optou-se por utilizar a metodologia PICOC (população, intervenção, comparação, resultados e contexto) proposta por Kitchenham (2007) para auxiliar na criação da *string* de busca. A pesquisa segue a estrutura que será apresentada a seguir.

3.1.1.1 Formulação de pesquisa

A pesquisa foi realizada seguindo a formulação de pesquisa:

a) Foco da questão

Esta revisão sistemática busca encontrar artigos que abordam técnicas de *machine learning* e *data mining*, visando verificar na literatura técnicas de análise preditiva para serem aplicadas na área de reabilitação pulmonar.

b) Questões de interesse

- Técnicas utilizadas
- Abordagens
- Estudos já realizados
- Aplicações recorrentes

c) Palavras-chave

Análise preditiva. *Machine Learning. Data Mining.* Reabilitação Pulmonar. Doenças respiratórias crônicas.

d) Intervenção

Identificar técnicas de análise preditiva que são utilizadas em programas de reabilitação pulmonar.

e) Controle

Não será utilizado.

f) Efeito

Verificar as oportunidades de pesquisa na área de inteligência artificial relacionadas à aplicação de análise preditiva em programas de reabilitação pulmonar.

g) Medida de resultado

Gerar embasamento teórico para o trabalho de conclusão de curso e para a escrita de artigos.

h) População de interesse

Pesquisadores, professores, desenvolvedores e profissionais da área da computação.

i) Aplicação

Esta pesquisa tem como foco pesquisadores, professores, desenvolvedores e profissionais da área da computação, pois visa identificar técnicas, abordagens e ferramentas para a aplicação de análise preditiva em programas de reabilitação pulmonar.

j) Desenho do experimento

Não será desenvolvido

k) Financiamento

Sem financiamento

3.1.1.2 Formulação de critérios

Os critérios da pesquisa foram definidos seguindo a estrutura:

a) Definição de critérios de seleção das fontes de dados

As fontes de dados utilizadas nesta revisão foram selecionadas por indicação dos orientadores desse projeto e através de trabalhos relacionados. Para a seleção das fontes de dados, foi observado se a fonte possui publicações nas áreas da computação e saúde. Será utilizada a base de dados *Web of Science* (WEB OF SCIENCE, 2018) para buscar trabalhos relacionados à área da computação. Essa base possui periódicos de assuntos diversos e seu classificador de periódicos apresenta maior precisão que outras bases de dados (FRANCESCHET, 2010) (VIEIRA; WAINER, 2013) (WANG; WALTMAN, 2016). Nesta base de dados, somente pesquisadores CAPES possuem acesso integral aos trabalhos. Em relação à área da saúde, será utilizada a base de dados *Pubmed* (PUBMED, 2018), que contém periódicos publicados nessa área (MEDEIROS, 2016). Essa base de dados possui periódicos do *Medline (Medical Literature Analysis and Retrieval System Online)*, possivelmente o banco de dados bibliográfico mais importante do mundo em ciências da vida e informações biomédicas (EDHLUND; MCDOUGALL, 2014).

b) Idiomas das fontes de dados

Somente serão consideradas as publicações que estiverem no idioma inglês.

c) *String* de busca

A *string* de busca foi definida para identificar somente trabalhos que possuam uma relação entre computação e saúde, visando buscar trabalhos mais próximos do objetivo desta revisão, que é a aplicação de técnicas de análise preditiva em programas de reabilitação pulmonar. Para auxiliar na definição dos termos relacionados à área da saúde, foi utilizado o DeCS (DECS, 2018), um vocabulário estruturado e trilíngue que foi desenvolvido para servir como uma linguagem única na indexação de artigos.

(("machine learning" or "predictive analysis" or "data mining" or "predictive modeling") and ("pulmonary rehabilitation" or "lung rehabilitation" or "pulmonary injur" or "lung injur*" or "respiratory disease" or "pulmonary function" or "lung function" or "cardiopulmonary disease" or "pulmonary disease" or "lung disease"))*

d) Artigos de controle

Optou-se por não utilizar nenhum artigo de controle para esta revisão sistemática.

3.1.1.3 Seleção dos estudos

Os estudos foram selecionados conforme a seguinte definição:

i. Critérios para inclusão/exclusão dos resultados

a) O ano de publicação do artigo deve estar dentro do período de 2014 a 2018;

b) Ser um artigo publicado;

c) Ser um artigo escrito em inglês;

d) A publicação deve estar disponível na íntegra na internet ou disponível através de convênios das instituições de ensino;

e) Ser validado pelo software *Tree of Science* (TOS) (GIRALDO; ZULUAGA; ESPINOSA, 2014);

f) O artigo deve possuir relação entre assuntos obrigatórios da pesquisa.

ii. Procedimentos para seleção dos estudos

Primeiramente, foram realizadas as consultas às bases de dados, executando a *string* de busca em cada uma das bases. Após isso, os trabalhos selecionados foram exportados das bases de dados e importados na ferramenta StArt (LAPES, 2018). Após isso, foram executadas as 5 fases definidas a seguir.

iii. Fases de seleção de artigos

Fase 1 – Validar os critérios de inclusão/exclusão;

Fase 2 – Validar as publicações da base de dados *Web of Science* na ferramenta TOS e validar os artigos duplicados;

Fase 3 – Leitura do título, palavras-chave e resumo;

Fase 4 – Leitura da introdução e conclusão;

Fase 5 – Leitura integral dos artigos e validação das respostas para os critérios de qualidade.

iv. Critérios de qualidade

- Quais foram os algoritmos mais utilizados?
- Quais foram as técnicas mais utilizadas?
- Os artigos foram aplicados à reabilitação pulmonar?
- Os artigos utilizaram alguma forma de validação?
- Quais foram as linguagens de programação mais utilizadas?
- A proposta teve resultados positivos?
- Quais foram as ferramentas mais utilizadas?

3.2 DESENVOLVIMENTO DA REVISÃO SISTEMÁTICA

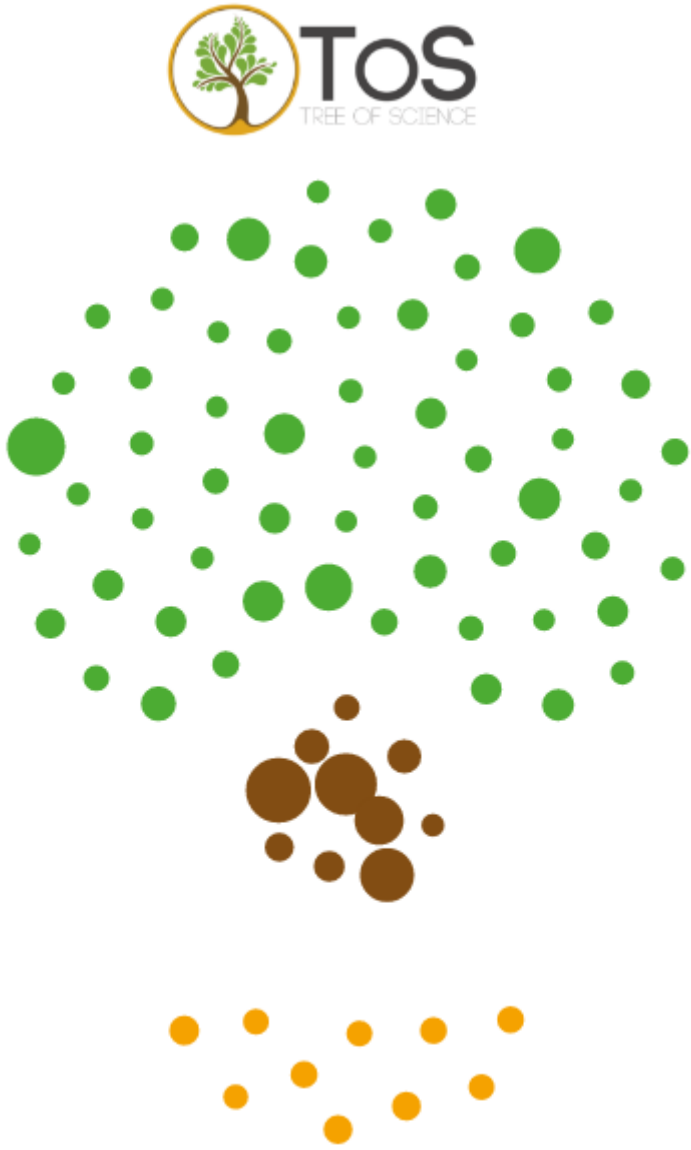
O desenvolvimento da revisão sistemática foi conduzido com o auxílio da ferramenta StArt. Esta ferramenta foi desenvolvida para auxiliar os pesquisadores na aplicação da técnica de revisão sistemática, desde o cadastro do protocolo até a fase final da revisão. Também são gerados gráficos com informações sobre o andamento do trabalho.

Primeiramente, foi cadastrado o protocolo na ferramenta. Após isso, foram importados os artigos exportados das bases de dados para início da seleção dos trabalhos. Por fim, foram executadas as fases definidas no protocolo citado anteriormente.

Os artigos identificados na base de dados Web of Science foram validados no software TOS. Esta ferramenta auxilia na análise dos artigos, selecionando os mais relevantes através da teoria de grafos, separando em três classes e representando em forma de árvore. A primeira classe são as folhas (destacadas em verde), que representam os trabalhos que são tendências sobre o assunto pesquisado. A segunda classe é o tronco (destacado em marrom), representam os trabalhos que formam a estrutura da pesquisa. A terceira classe é a raiz (destacada em laranja), que são os trabalhos primários em relação ao assunto pesquisado. Nesta revisão, os artigos da raiz não foram utilizados por terem sido publicados antes do ano de 2014, não

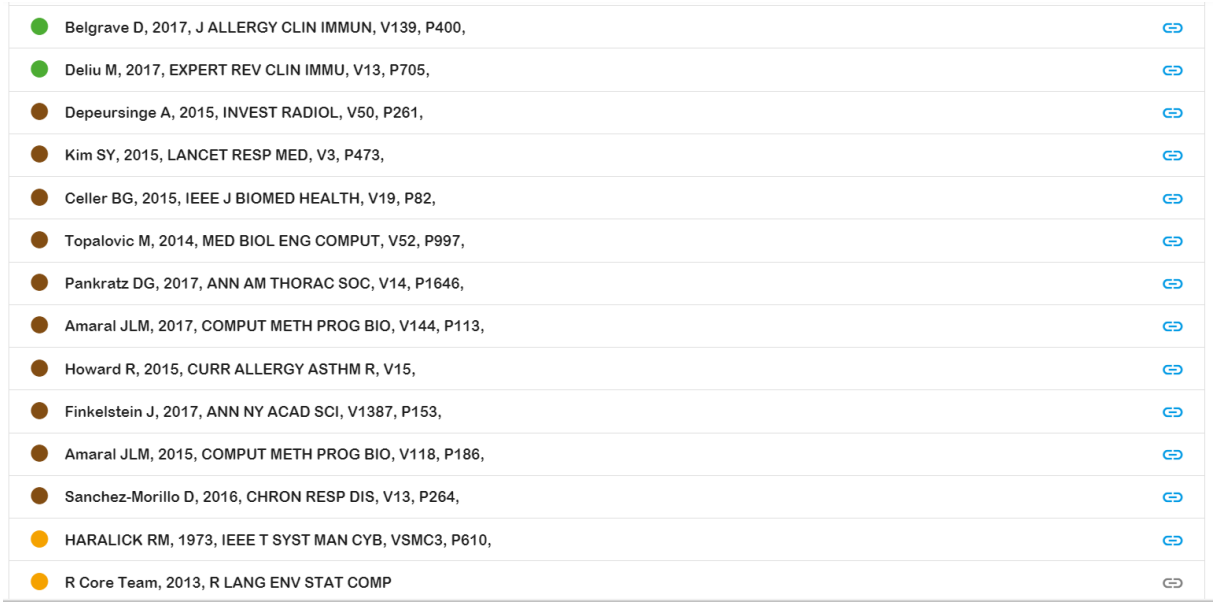
atendendo ao critério de inclusão referente ao período de publicação. Os artigos da base de dados Pubmed não são compatíveis com o TOS e, por isso, não foram validados nessa ferramenta. A Figura 3 demonstra a árvore gerada pela ferramenta TOS a partir da importação dos trabalhos identificados. A Figura 4 apresenta alguns artigos que formam a árvore gerada pela ferramenta TOS.

Figura 3 - Árvore gerada pela ferramenta TOS.



Fonte: elaborado pelo autor

Figura 4 – Demonstração de artigos que formam a árvore gerada pela ferramenta TOS.

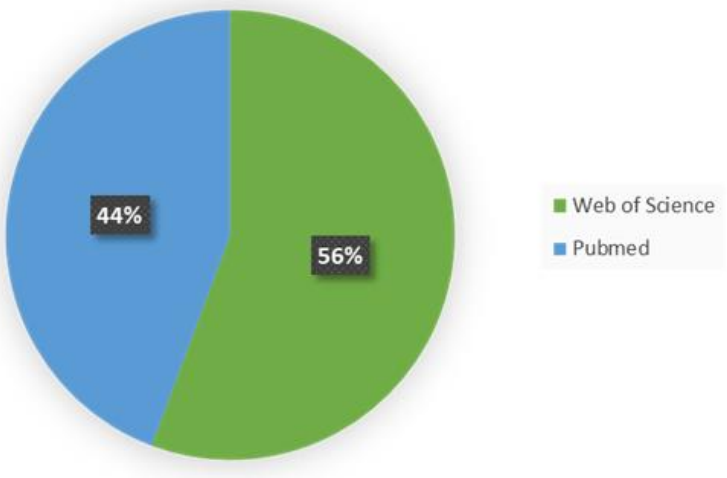


Fonte: elaborado pelo autor

3.2.1 Fases de seleção

A consulta às bases de dados com a *string* de busca definida anteriormente foi realizada no dia 12 de abril de 2018. Na base de dados *Web of Science*, foram identificados 181 artigos. Na base de dados *Pubmed*, foram identificados 144 artigos. Ao todo, foram selecionados 325 artigos para o início desta revisão sistemática. A Figura 5 demonstra o gráfico comparativo de resultados entre as bases de dados.

Figura 5 - Gráfico comparativo de resultados entre as bases selecionadas.



Fonte: elaborado pelo autor

3.2.1.1 Fase 1

A primeira fase consistiu na validação dos critérios de inclusão/exclusão. Nesta fase, foram removidos os artigos publicados anteriormente a 2014 e os artigos com o idioma diferente de inglês. Os critérios foram aplicados diretamente nas bases de dados, visto que os motores de busca possibilitam diversos filtros. Após essa fase, os artigos foram exportados das bases de dados e importados na ferramenta StArt. Ao final da primeira fase, 223 artigos foram aceitos para a próxima fase, sendo 120 da base *Web of Science* e 103 da base *Pubmed*.

3.2.1.2 Fase 2

Na segunda fase, foram validados os artigos na ferramenta TOS. Conforme mencionado anteriormente, somente os artigos da base de dados *Web of Science* foram validados pelo fato da ferramenta não ser compatível com a base de dados *Pubmed*. Além disso, foram removidos os artigos duplicados.

Ao validar os artigos na ferramenta TOS, foram selecionados 80 artigos. Desses, os 10 artigos classificados como raiz da árvore gerada não foram utilizados por terem sido publicados antes de 2014. Desta forma, 50 artigos da base *Web of Science* foram rejeitados. Nesta fase, também foi realizada uma avaliação manual de artigos duplicados, resultando na classificação de 57 artigos como duplicados. No final desta fase, foram aceitos 116 artigos para a próxima fase.

3.2.1.3 Fase 3

Na terceira fase, foram lidos título, palavras-chave e resumo para avaliar os artigos selecionados na fase anterior. Foram removidos 96 artigos por não possuírem relação entre dois assuntos obrigatórios da pesquisa. Ao final desta fase, foram aceitos 20 artigos para a quarta fase.

3.2.1.4 Fase 4

Nesta fase, foram lidos introdução e conclusão de todos os artigos selecionados na terceira fase. Os artigos foram analisados para avaliar se estavam de acordo com os assuntos obrigatórios relacionados à esta revisão. Nenhum artigo foi removido nesta fase, resultando em 20 artigos aceitos para a última fase.

3.2.1.5 Fase 5

A última fase desta revisão consistiu na leitura integral dos 20 artigos selecionados. Juntamente a isso, foram validados os critérios de qualidade definidos no protocolo desta revisão. Ao final desta fase, foi iniciada a análise das informações extraídas dos artigos. A Figura 6 apresenta o progresso das fases desta revisão sistemática. É apresentada a quantidade de artigos selecionados no final de cada fase e o percentual de redução de artigos em relação à fase anterior.

Figura 6 - Progresso da revisão sistemática

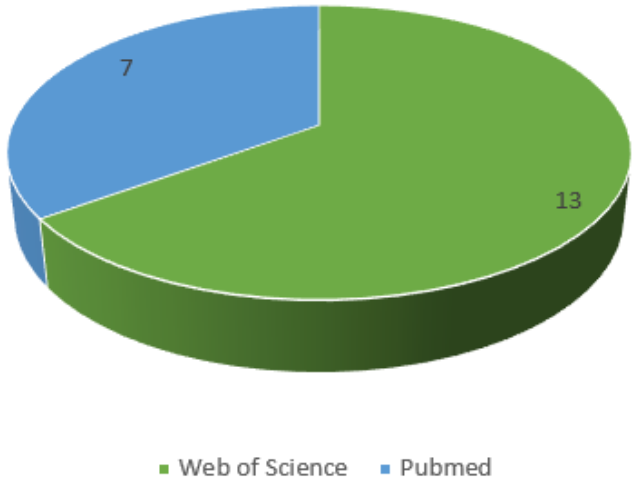
FASE INICIAL		FASE 1		FASE 2		FASE 3		FASE 4	
Base	Artigos	Artigos	Redução	Artigos	Redução	Artigos	Redução	Artigos	Redução
Web of Science	181	120	33,70%	70	41,67%	13	81,43%	13	0,00%
Pubmed	144	103	28,47%	46	55,34%	7	84,78%	7	0,00%
Total	325	223	31,38%	116	47,98%	20	82,76%	20	0,00%

Fonte: elaborado pelo autor

3.3 RESULTADOS

Após a leitura completa dos artigos, foram respondidas as questões referentes aos critérios de qualidade estabelecidos no protocolo. Os resultados serão debatidos nesta seção. O Quadro 1 apresenta os artigos selecionados para a extração de dados.

Figura 7 - Gráfico comparativo de resultados entre as bases selecionadas.



Fonte: elaborado pelo autor

Quadro 1 - Quadro de artigos selecionados.

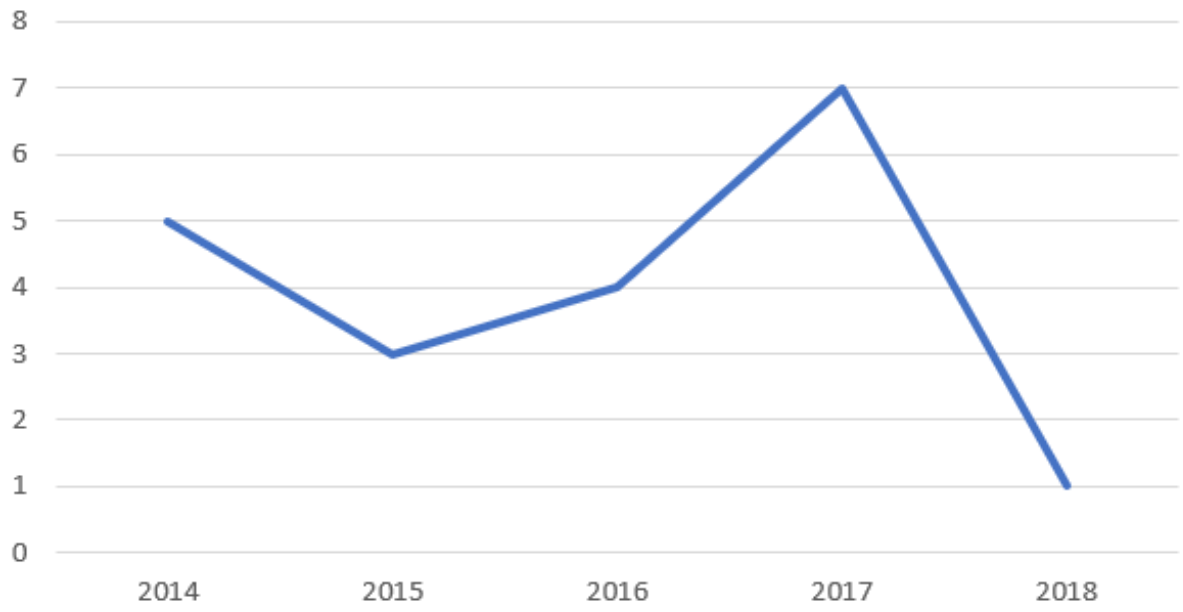
Base de Dados	Título	Ano
Web of Science	A Machine Learning Approach to Prediction of Exacerbations of Chronic Obstructive Pulmonary Disease	2015
Web of Science	A machine learning approach to triaging patients with chronic obstructive pulmonary disease	2017
Web of Science	AN EMPIRICAL STUDY OF SKEW-INSENSITIVE SPLITTING CRITERIA AND ITS APPLICATION IN TRADITIONAL CHINESE MEDICINE	2014
Pubmed	Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential	2018
Web of Science	A systematic review of predictive models for asthma development in children	2015
Web of Science	Automated Interpretation of Pulmonary Function Tests in Adults with Respiratory Complaints	2017
Pubmed	Diagnosing asthma and chronic obstructive pulmonary disease with machine learning.	2017
Web of Science	Exacerbations in Chronic Obstructive Pulmonary Disease: Identification and Prediction Using a Digital Health System	2017
Web of Science	High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements	2017
Pubmed	Insight into Best Variables for COPD Case Identification: A Random Forests Analysis.	2016
Pubmed	Interpretable Deep Models for ICU Outcome Prediction.	2016
Web of Science	Learning Bayesian networks for clinical time series analysis	2014
Web of Science	Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease	2015
Web of Science	Machine learning approaches to personalize early prediction of asthma exacerbations	2017
Web of Science	Modelling the dynamics of expiratory airflow to describe chronic obstructive pulmonary disease	2014
Pubmed	Reducing COPD readmissions through predictive modeling and incentive-based interventions.	2017
Pubmed	Text mining and medicine: usefulness in respiratory diseases.	2014
Pubmed	Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics.	2016
Web of Science	Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data	2014
Web of Science	Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review	2016

Fonte: elaborado pelo autor

Foram gerados alguns gráficos para facilitar a análise e visualização de informações sobre os artigos selecionados. A Figura 7 demonstra a distribuição dos artigos selecionados conforme a base de dados. Pode-se observar que a maior parte dos artigos é oriundo da base de dados *Web of Science*.

A Figura 8 demonstra a distribuição dos artigos conforme o ano de publicação. Pode-se observar um crescimento de trabalhos publicados na área nos últimos anos. É importante ressaltar que, em relação ao ano de 2018, somente foram considerados os trabalhos publicados até 12 de abril, data em que foi realizada a pesquisa nas bases de dados. Desta forma, os trabalhos publicados após essa data não foram considerados para esta revisão.

Figura 8 - Gráfico comparativo de publicações por ano.



Fonte: elaborado pelo autor

3.3.1 Perguntas respondidas

Após a leitura completa dos artigos, foram validados os critérios de qualidade estabelecidos no protocolo. As respostas para as 7 perguntas são demonstradas nos quadros a seguir. Todas as perguntas possibilitam mais de uma resposta com exceção das perguntas 3 e 6.

1. Quais foram os algoritmos mais utilizados?

Quadro 2 - Algoritmos mais utilizados.

Algoritmo	Artigos
Não informado	11
K-Nearest Neighbour Classifier (KNN)	8
C4.5	1
Hellinger Distance Decision Tree	1
Bhattacharyya Distance Decision Tree	1
Chi-squared Divergence Decision Tree	1
Kullback-Leibler Divergence Decision Tree	1
AdaBoost	1

Fonte: elaborado pelo autor

2. Quais foram as técnicas mais utilizadas?

Quadro 3 - Técnicas mais utilizadas.

Técnica	Artigos
Support Vector Machine	7
Decision Tree	6
Logistic Regression	6
Random Forest	5
Naive Bayes	5
Artificial Neural Networks (ANN)	3
Bayesian Network	3
Radial Basis Function Neural Network (RBF)	2
K-Means Classifier Clustering	2
Probabilistic Neural Network (PNN)	2
Least Squares Support Vector Machine (LS-SVM)	2
Gradient Boosted Random Forest	1
Extra Decision Tree Classifier	1
Convolutional Neural Network (CNN)	1
Dissimilarity based Partial Least Squares (DPLS)	1
Feature-Based Dissimilarity Space Classifier (FDSC)	1
Gradient Boosting Trees (GBT)	1
Linear Bayes Normal Classifier (LBNC)	1
Text Mining	1
Linear Discriminant Analysis (LDA)	1
Não informado	1

Fonte: elaborado pelo autor

3. Os algoritmos foram aplicados à reabilitação pulmonar?

Quadro 4 - Aplicação em reabilitação pulmonar.

Aplicação em reabilitação pulmonar	Artigos
Não	20

Fonte: elaborado pelo autor

4. Os artigos utilizaram alguma forma de validação?

Quadro 5 - Formas de validação mais utilizadas.

Validação	Artigos
Sensibilidade e Especificidade	11
Area Under the ROC Curve (AUC)	10
10-fold cross validation	6
5-fold cross validation	4
Não informado	4
k-fold cross validation	3
Painel de médicos	2
2-fold cross validation	2
Bootstrapping	2
Area Under Precision-Recall Curve (AUPRC)	1
Escore de Brier	1

Fonte: elaborado pelo autor

5. Quais foram as linguagens de programação mais utilizadas?

Quadro 6 - Linguagens de programação mais utilizadas.

Linguagem de programação	Artigos
Não informado	14
R	5
Python	2

Fonte: elaborado pelo autor

6. A proposta teve resultados positivos?

Quadro 7 - Resultados.

Resultados Positivos	Artigos
Sim	14
Não	1
Não possui resultados	1
Sem aplicação	4

Fonte: elaborado pelo autor

7. Quais foram as ferramentas mais utilizadas?

Quadro 8 - Ferramentas mais utilizadas.

Ferramenta/*Framework*	**Artigos**
Não informado	12
Pacote Python Scikit-Learn	2
Pacote R optFederov	1
Pacote R randomForest	1
Pacote R boot	1
Oracle Data Miner	1
PRTools	1
LS-SVMlab	1
Pacote R glmnet	1
Pacote R pamr	1
SPSS	1

Fonte: elaborado pelo autor

3.3.2 Análise dos artigos

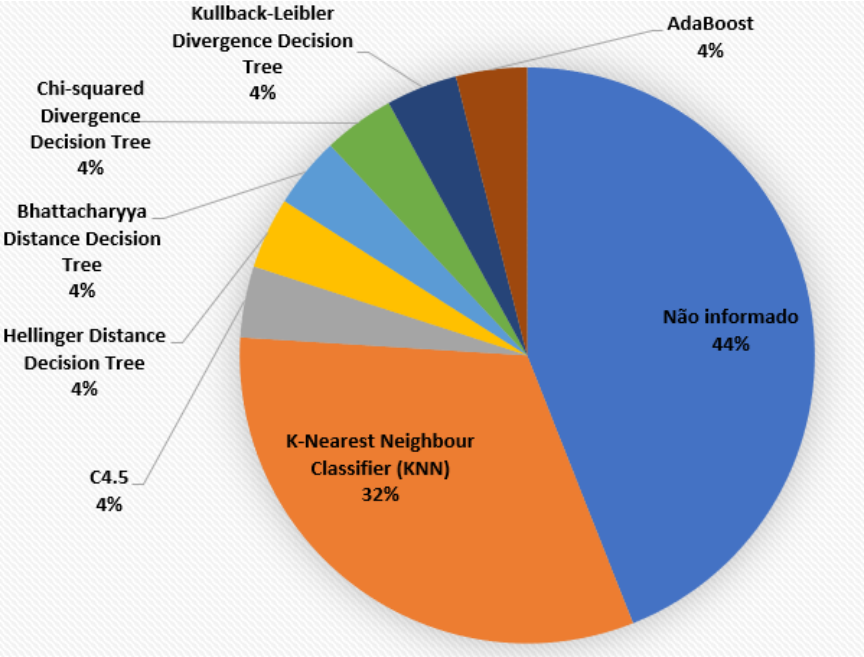
A partir da validação dos critérios de qualidade, foi realizada uma análise detalhada sobre os artigos selecionados. Nesta análise, foram gerados gráficos quantitativos e observadas as características dos artigos. A análise detalhada será apresentada nesta seção.

3.3.2.1 Algoritmos

Após a leitura dos artigos, pode-se observar que a maior parte dos trabalhos não especifica o algoritmo utilizado, totalizando 11 trabalhos. Além disso, pode-se observar que o algoritmo KNN se destacou entre os trabalhos que especificaram o algoritmo utilizado, sendo mencionado em 8 trabalhos (SWAMINATHAN *et al.*, 2017) (DAS; TOPALOVIC; JANSSENS, 2017) (SPATHIS; VLAMOS, 2017) (AMARAL *et al.*, 2017) (AMARAL *et al.*, 2015) (TOPALOVIC *et al.*, 2014) (GUAN *et al.*, 2016) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016). Desconsiderando as revisões sistemáticas, foram identificados dois trabalhos que utilizaram mais de um algoritmo para comparação de desempenho. O primeiro (SU *et al.*, 2014) comparou os algoritmos C4.5, *Hellinger Distance Decision Tree, Bhattacharyya Distance Decision Tree, Chi-squared Divergence Decision Tree* e *Kullback-Leibler Divergence Decision Tree*, sendo o *Kullback-Leibler Divergence Decision Tree* o algoritmo com melhor desempenho. Esse foi o único trabalho que utilizou essas técnicas. O segundo (AMARAL *et al.*, 2017) utilizou os algoritmos *AdaBoost* e KNN, relatando desempenho semelhante entre ambos. Esse foi o único

trabalho que utilizou o algoritmo *AdaBoost*. A Figura 9 demonstra o gráfico comparativo de algoritmos utilizados nos trabalhos selecionados.

Figura 9 - Gráfico comparativo de algoritmos utilizados.



Fonte: elaborado pelo autor

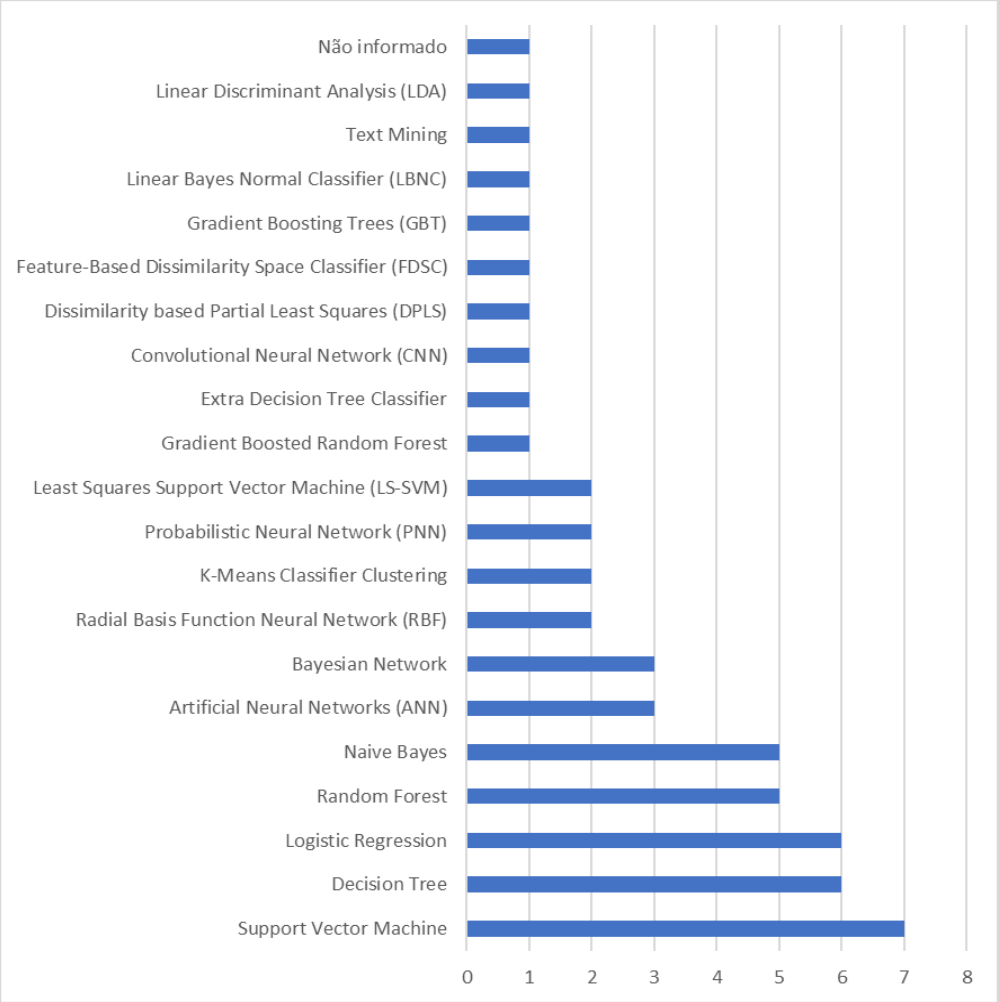
3.3.2.2 Técnicas

Os trabalhos selecionados utilizaram diversas técnicas para comparação de resultados. Apenas em um trabalho não foi informada a técnica utilizada. A Figura 10 apresenta o gráfico comparativo de técnicas utilizadas nos artigos.

Pode-se observar que a técnica *Support Vector Machine* foi a mais utilizada, aparecendo em 7 trabalhos (SWAMINATHAN *et al.*, 2017) (DAS; TOPALOVIC; JANSSENS, 2017) (SPATHIS; VLAMOS, 2017) (AMARAL *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (WU *et al.*, 2014) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016). Outras técnicas que se destacaram foram *Decision Tree* (SU *et al.*, 2014) (TOPALOVIC *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (AMARAL *et al.*, 2017) (TOPALOVIC *et al.*, 2014) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016) e *Logistic Regression* (SWAMINATHAN *et al.*, 2017) (LUO *et al.*, 2015) (SPATHIS; VLAMOS, 2017) (SHAH *et al.*, 2017) (ZHONG *et al.*, 2017) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016), que foram utilizadas em 6 trabalhos cada uma. Na sequência, *Random Forest* (SWAMINATHAN *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (AMARAL *et al.*, 2017) (LEIDY *et al.*, 2016) (AMARAL *et*

al., 2015) e *Naive Bayes* (SWAMINATHAN *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (HEIJDEN; VELIKOVA; LUCAS, 2014) (FINKELSTEIN; JEONG, 2017) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016) foram citadas em 5 trabalhos cada uma.

Figura 10 - Gráfico comparativo de técnicas utilizadas.



Fonte: elaborado pelo autor

Foram identificados 8 trabalhos (FERNANDEZ-GRANERO *et al.*, 2015) (SWAMINATHAN *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (AMARAL *et al.*, 2017) (HEIJDEN; VELIKOVA; LUCAS, 2014) (AMARAL *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (TOPALOVIC *et al.*, 2014) que utilizaram mais de uma técnica para comparação de resultados. Nessa contagem, as revisões sistemáticas foram desconsideradas, visto que não é realizada uma comparação de desempenho das técnicas abordadas nesses trabalhos. Em geral, não foi identificada uma regra para seleção de técnicas a fim de comparação. Cada trabalho comparou técnicas variadas para medir seus desempenhos. Em contrapartida, foram

identificados 7 trabalhos que utilizaram apenas uma técnica. Nesses trabalhos, foram aplicadas as técnicas *Decision Tree* (SU *et al.*, 2014) (TOPALOVIC *et al.*, 2017), *Logistic Regression* (SHAH *et al.*, 2017) (ZHONG *et al.*, 2017), *Random Forest* (LEIDY *et al.*, 2016), *Gradient Boosting Trees* (GBT) (CHE *et al.*, 2016) e *Support Vector Machine* (WU *et al.*, 2014).

3.3.2.3 Aplicação na área de reabilitação pulmonar

Na seleção de artigos para esta revisão sistemática, não foi localizada nenhuma proposta ou revisão sistemática que abordou a aplicação de análise preditiva na área de reabilitação pulmonar. Os artigos selecionados abordam, em geral, o diagnóstico (SWAMINATHAN *et al.*, 2017) (SU *et al.*, 2014) (DAS; TOPALOVIC; JANSSENS, 2017) (SPATHIS; VLAMOS, 2017) (LEIDY *et al.*, 2016), o grau de obstrução de vias aéreas (AMARAL *et al.*, 2017) (AMARAL *et al.*, 2015), fenótipos (GUAN *et al.*, 2016) (WU *et al.*, 2014) e exacerbações (FERNANDEZ-GRANERO *et al.*, 2015) (SHAH *et al.*, 2017) (FINKELSTEIN; JEONG, 2017) de doenças respiratórias crônicas. Desta forma, pode-se observar a existência de oportunidades de pesquisa relacionadas à aplicação de técnicas de análise preditiva na área de reabilitação pulmonar.

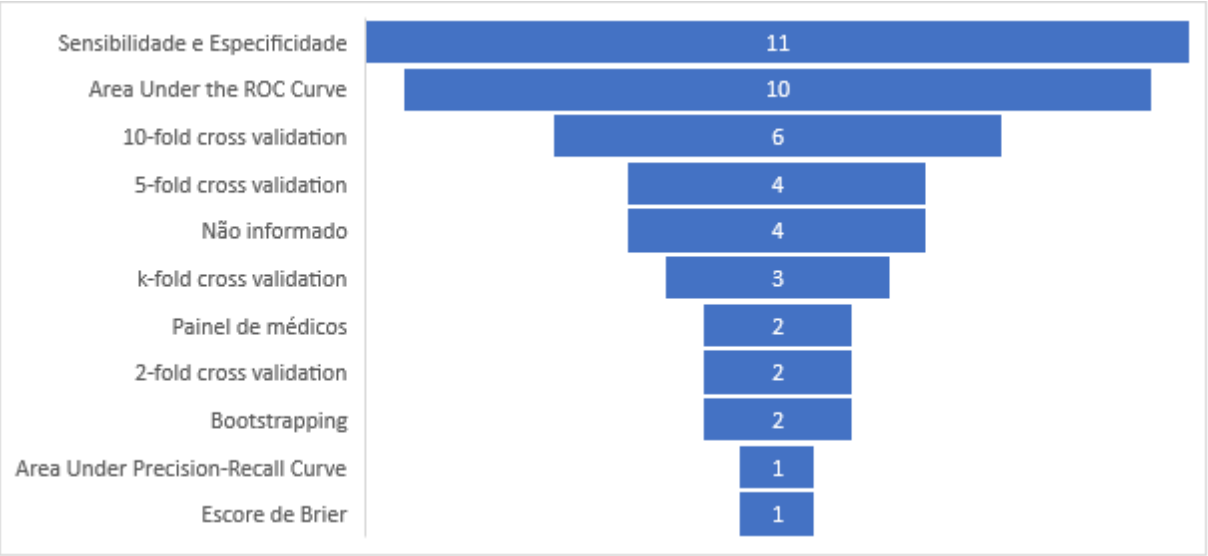
3.3.2.4 Validação

Em relação à validação das propostas, observou-se que foram utilizadas diversas formas de validação. O gráfico apresentado na Figura 11 demonstra uma análise quantitativa das formas de validação utilizadas nos trabalhos. Destacaram-se sensibilidade e especificidade, seguidos por *Area Under the Receiver Operating Characteristic Curve* (AUC) e validações realizadas diretamente nas técnicas de análise preditiva, tais como *10-fold cross validation* e *5-fold cross validation.*

É possível avaliar que grande parte dos trabalhos utilizaram as métricas sensibilidade e especificidade para validar a proposta (FERNANDEZ-GRANERO *et al.*, 2015) (SWAMINATHAN *et al.*, 2017) (LUO *et al.*, 2015) (TOPALOVIC *et al.*, 2017) (SHAH *et al.*, 2017) (AMARAL *et al.*, 2017) (LEIDY *et al.*, 2016) (AMARAL *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (TOPALOVIC *et al.*, 2014) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016). Segundo Ferreira e Patino (2017), esses indicadores são excelentes para avaliar o desempenho de um novo teste diagnóstico de doenças em comparação com um teste já estabelecido. A sensibilidade indica a proporção de

diagnósticos corretos da doença em indivíduos que possuem a doença. Já a especificidade é a proporção de diagnósticos corretos em indivíduos que não possuem a doença.

Figura 11 - Gráfico comparativo de formas de validação utilizadas.



Fonte: elaborado pelo autor

Ao todo, 14 trabalhos utilizaram validações na própria técnica (FERNANDEZ-GRANERO et al., 2015) (SWAMINATHAN et al., 2017) (SU et al., 2014) (TOPALOVIC et al., 2017) (SPATHIS; VLAMOS, 2017) (SHAH et al., 2017) (AMARAL et al., 2017) (CHE et al., 2016) (AMARAL et al., 2015) (FINKELSTEIN; JEONG, 2017) (TOPALOVIC et al., 2014) (ZHONG et al., 2017) (WU et al., 2014) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016). Apenas 2 trabalhos validaram a proposta através de especialistas na área (SWAMINATHAN et al., 2017) (TOPALOVIC et al., 2017). Esses 2 trabalhos também utilizaram uma validação na técnica juntamente à sensibilidade e especificidade, comparando o resultado com a validação de especialistas. Desconsiderando as revisões sistemáticas, foram identificados 12 trabalhos (FERNANDEZ-GRANERO et al., 2015) (SWAMINATHAN et al., 2017) (TOPALOVIC et al., 2017) (SPATHIS; VLAMOS, 2017) (SHAH et al., 2017) (AMARAL et al., 2017) (CHE et al., 2016) (HEIJDEN; VELIKOVA; LUCAS, 2014) (AMARAL et al., 2015) (FINKELSTEIN; JEONG, 2017) (TOPALOVIC et al., 2014) (ZHONG et al., 2017) que compararam mais de uma técnica de validação, sejam elas diretamente na técnica, através de especialistas ou de outros indicadores. Nas validações, também não foi identificada uma regra para seleção de formas a fim de comparação. Os trabalhos compararam diversas formas de validação para medir seus desempenhos.

3.3.2.5 Linguagens de programação

Grande parte dos artigos não especificou a linguagem de programação utilizada. Dentre os trabalhos que especificaram, a linguagem R foi a que se destacou, sendo utilizada em 5 trabalhos (SWAMINATHAN *et al.*, 2017) (LEIDY *et al.*, 2016) (HEIJDEN; VELIKOVA; LUCAS, 2014) (ZHONG *et al.*, 2017) (GUAN *et al.*, 2016). Outra linguagem mencionada foi a linguagem Pyhton. Essa linguagem foi utilizada em 2 trabalhos (SWAMINATHAN *et al.*, 2017) (SPATHIS; VLAMOS, 2017).

3.3.2.6 Resultados

A maior parte dos artigos apresentou e aplicou uma proposta, gerando, desta forma, resultados ao final da aplicação. Após analisar os artigos, observou-se que 14 trabalhos (FERNANDEZ-GRANERO *et al.*, 2015) (SWAMINATHAN *et al.*, 2017) (SU *et al.*, 2014) (TOPALOVIC *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (SHAH *et al.*, 2017) (AMARAL *et al.*, 2017) (LEIDY *et al.*, 2016) (CHE *et al.*, 2016) (HEIJDEN; VELIKOVA; LUCAS, 2014) (AMARAL *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (TOPALOVIC *et al.*, 2014) (WU *et al.*, 2014) obtiveram resultados positivos a partir da aplicação da proposta e apenas um trabalho (GUAN *et al.*, 2016) apresentou resultados negativos. No entanto, alguns autores (FERNANDEZ-GRANERO *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (AMARAL *et al.*, 2017) (SPATHIS; VLAMOS, 2017) utilizaram uma amostragem relativamente pequena para a aplicação da proposta e observaram que os resultados podem ter sido afetados por isso. Um trabalho não obteve resultados (ZHONG *et al.*, 2017), pois ainda não aplicou de forma prática o modelo apresentado, abordando apenas o conceito teórico da proposta. Dentre os trabalhos, também existem revisões sistemáticas (DAS; TOPALOVIC; JANSSENS, 2017) (LUO *et al.*, 2015) (PIEDRA; FERRER; GEA, 2014) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016), que mencionaram outros trabalhos na área que obtiveram resultados positivos em geral.

3.3.2.7 Ferramentas/frameworks

Assim como a maior parte dos trabalhos não apresentou a linguagem de programação utilizada, também foi possível observar que grande parte dos trabalhos não citou a utilização de ferramentas para auxiliar na aplicação da proposta. Desta forma, 12 trabalhos não informaram a utilização de ferramentas auxiliares. Dos artigos que informaram a utilização de ferramentas, o Pacote Python “Scikit-Learn” foi destaque, aparecendo em dois trabalhos

(SWAMINATHAN et al., 2017) (SPATHIS; VLAMOS, 2017). Desta forma, os dois trabalhos que utilizaram a linguagem de programação Python também utilizaram o Pacote "Scikit-Learn". Em relação à linguagem de programação R, foram utilizados os pacotes "optFederov" (SWAMINATHAN et al., 2017), "randomForest" (LEIDY et al., 2016), "boot" (HEIJDEN; VELIKOVA; LUCAS, 2014), "glmnet" (ZHONG et al., 2017) e "pamr" (GUAN et al., 2016), presentes em um trabalho cada. Também foram utilizadas, em um trabalho cada, as ferramentas "Oracle Data Miner" (FINKELSTEIN; JEONG, 2017), "PRTools" (TOPALOVIC et al., 2014), "LS-SVMlab" (TOPALOVIC et al., 2014) e "SPSS" (GUAN et al., 2016). Acredita-se que não foi identificada uma ferramenta que se destacou com muitas citações nos trabalhos devido à variedade de ferramentas disponíveis atualmente e à diversidade de propostas, técnicas e linguagens de programação utilizadas nos trabalhos selecionados.

3.4 ANÁLISE CRÍTICA

A utilização de inteligência artificial através da aplicação de técnicas de análise preditiva pode gerar grandes benefícios para a área da medicina pulmonar. Essas técnicas podem ser adaptadas para problemas variados dessa área, auxiliando em processos de cuidados com a saúde. Atualmente, existem diversas aplicações dessas técnicas na área da saúde e o número de estudos na área vem crescendo nos últimos anos. No entanto, é possível identificar que existem algumas áreas que ainda não foram exploradas, como, por exemplo, a utilização de análise preditiva em programas de reabilitação pulmonar.

Os estudos analisados demonstram que a utilização da análise preditiva na área de medicina pulmonar é diversificada. Dentre as principais utilizações estão a predição de exacerbações (FERNANDEZ-GRANERO *et al.*, 2015) (SHAH *et al.*, 2017) (FINKELSTEIN; JEONG, 2017), desenvolvimento (LUO *et al.*, 2015) e diagnóstico de doenças respiratórias crônicas (SWAMINATHAN *et al.*, 2017) (SU *et al.*, 2014) (SPATHIS; VLAMOS, 2017) (LEIDY *et al.*, 2016). Outras aplicações identificadas foram a detecção de obstrução de vias aéreas (AMARAL *et al.*, 2017) (AMARAL *et al.*, 2015), a identificação de fenótipos de doenças respiratórias (GUAN *et al.*, 2016) (WU *et al.*, 2014), a análise de som pulmonar para detecção de doenças, a análise de dados de sistemas de telemedicina, a análise de testes de função pulmonar e a análise de dados de tomografia computadorizada (DAS; TOPALOVIC; JANSSENS, 2017).

A análise preditiva parte da construção de um modelo preditivo. É possível afirmar que, para atingir os resultados desejados, esse modelo deve atender a dois pontos importantes, ter bons preditores e utilizar uma técnica adequada de análise preditiva. Os bons preditores resultam em boas inferências que podem ser realizadas sobre os dados. Uma técnica adequada auxilia tanto no desempenho preditivo quanto no desempenho computacional. Ambos os fatores serão debatidos a seguir.

3.4.1 Preditores

Vários autores adotaram como primeira etapa da construção do modelo preditivo o estudo de preditores para o problema abordado (SWAMINATHAN *et al.*, 2017) (FINKELSTEIN; JEONG, 2017) (WU *et al.*, 2014) (LEIDY *et al.*, 2016) (GUAN *et al.*, 2016) (ZHONG *et al.*, 2017). Essa seleção de preditores foi realizada por meio de trabalhos relacionados à área estudada ou pela identificação, a partir de testes de aplicação, de uma técnica em uma base de dados de teste. Por consequência da boa seleção de preditores, os resultados obtidos são melhores.

Os preditores estão bastante ligados ao problema estudado. Nos trabalhos que tiveram como objetivo o diagnóstico de doenças respiratórias crônicas, por exemplo, um trabalho utilizou fatores já conhecidos na medicina por serem indicadores ou fatores de risco da doença em questão para obter melhores resultados (SWAMINATHAN *et al.*, 2017). Já em relação à identificação a partir de testes na base estudada, o resultado é dependente da diversidade da base de dados (SU *et al.*, 2014).

Em conjunto com a diversidade da base de dados estudada, o tamanho da amostragem e quantidade de preditores também pode influenciar nos resultados. Essas variáveis também estão diretamente ligadas ao problema estudado. No geral, os autores obtiveram resultados positivos nos trabalhos selecionados. No entanto, alguns autores (FERNANDEZ-GRANERO *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (AMARAL *et al.*, 2017) (SPATHIS; VLAMOS, 2017) relataram que o tamanho da amostragem pode ter impactado nos resultados do trabalho. Desta forma, a amostragem relativamente pequena foi citada como limitação de alguns trabalhos.

A identificação de bons preditores melhora o desempenho do modelo preditivo. Porém, o modelo não depende unicamente dos preditores. Para complementar a seleção de

bons preditores, a escolha de uma boa técnica também pode afetar o desempenho do modelo. As principais técnicas serão abordadas a seguir.

3.4.2 Técnicas

As técnicas são parte fundamental do modelo preditivo. A seleção da técnica também é dependente do problema estudado. No entanto, não foi identificada uma única técnica para aplicação em estudos da área. Observou-se, nesta revisão sistemática, que vários autores (FERNANDEZ-GRANERO *et al.*, 2015) (SWAMINATHAN *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (AMARAL *et al.*, 2017) (HEIJDEN; VELIKOVA; LUCAS, 2014) (AMARAL *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (TOPALOVIC *et al.*, 2014) não optaram por utilizar uma única técnica para a criação do modelo, mas sim por utilizar múltiplas técnicas para comparação de resultados.

Para análise das principais técnicas, foi elaborada uma tabela. Serão analisadas as 5 técnicas mais citadas e os trabalhos nos quais foram citadas. O Quadro 9 apresenta um comparativo de técnicas mais utilizadas nos trabalhos selecionados nesta revisão.

Após a análise dos trabalhos e das técnicas mais citadas, pode-se observar que 80% dos artigos citaram pelo menos uma dessas técnicas (SWAMINATHAN *et al.*, 2017) (SU *et al.*, 2014) (DAS; TOPALOVIC; JANSSENS, 2017) (LUO *et al.*, 2015) (TOPALOVIC *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (SHAH *et al.*, 2017) (AMARAL *et al.*, 2017) (LEIDY *et al.*, 2016) (HEIJDEN; VELIKOVA; LUCAS, 2014) (AMARAL *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (TOPALOVIC *et al.*, 2014) (ZHONG *et al.*, 2017) (WU *et al.*, 2014) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016). Desses, aproximadamente 37% compararam mais de uma técnica para atingir o melhor resultado (SWAMINATHAN *et al.*, 2017) (SPATHIS; VLAMOS, 2017) (AMARAL *et al.*, 2017) (AMARAL *et al.*, 2015) (FINKELSTEIN; JEONG, 2017) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016). Um trabalho (SPATHIS; VLAMOS, 2017) comparou o desempenho de todas as 5 técnicas e dois trabalhos (SWAMINATHAN *et al.*, 2017) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016) citaram a utilização de 4 das principais técnicas.

Legenda do Quadro 9:
SVM *Support Vector Machine*
DT *Decision Tree*

LR *Logistic Regression*

RF *Random Forest*

NB *Naive Bayes*

Quadro 9 - Quadro comparativo de técnicas mais utilizadas.

Artigo	SVM	DT	LR	RF	NB
A machine learning approach to triaging patients with chronic obstructive pulmonary disease	X		X	X	X
An Empirical Study of Skew-Insensitive Splitting Criteria and its Application in Traditional Chinese Medicine		X			
Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential	X				
A systematic review of predictive models for asthma development in children			X		
Automated Interpretation of Pulmonary Function Tests in Adults with Respiratory Complaints		X			
Diagnosing asthma and chronic obstructive pulmonary disease with machine learning	X	X	X	X	X
Exacerbations in Chronic Obstructive Pulmonary Disease: Identification and Prediction Using a Digital Health System			X		
High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements		X		X	
Insight into Best Variables for COPD Case Identification: A Random Forests Analysis				X	
Learning Bayesian networks for clinical time series analysis					X
Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease	X			X	
Machine learning approaches to personalize early prediction of asthma exacerbations	X				X
Modelling the dynamics of expiratory airflow to describe chronic obstructive pulmonary disease		X			
Reducing COPD readmissions through predictive modeling and incentive-based interventions			X		
Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data	X				
Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review	X	X	X		X

Fonte: elaborado pelo autor

O trabalho que comparou o desempenho de todas as técnicas (SPATHIS; VLAMOS, 2017) tinha como objetivo o diagnóstico de asma e doença pulmonar obstrutiva crônica e o classificador *Random Forest* obteve o melhor desempenho de predição para o diagnóstico de ambas as doenças. Outro trabalho (AMARAL *et al.*, 2015) também relatou que esse classificador obteve o melhor desempenho dentre as técnicas avaliadas para a categorização da gravidade da obstrução das vias aéreas na doença pulmonar obstrutiva crônica. A técnica *Logistic Regression* obteve o melhor desempenho em um trabalho que tinha como objetivo a triagem de pacientes com doença pulmonar obstrutiva crônica (SWAMINATHAN *et al.*, 2017).

Nesta revisão, foram identificadas revisões sistemáticas (LUO et al., 2015) (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016). Uma das revisões sistemáticas (LUO et al., 2015) aborda modelos preditivos para o desenvolvimento de asma em crianças. Essa revisão mencionou apenas o classificador *Logistic Regression* como técnica para a predição. Outra revisão sistemática (SANCHEZ-MORILLO; FERNANDEZ-GRANERO; LEON-JIMENEZ, 2016) abordou a utilização de algoritmos preditivos em sistemas de monitoramento domiciliar de doença pulmonar obstrutiva crônica e asma. Essa revisão citou o uso de 4 das principais técnicas, não mencionando apenas o classificador *Random Forest*.

Embora não tenham sido identificadas técnicas que foram aplicadas diretamente para a área de reabilitação pulmonar, as técnicas identificadas nesta revisão são bastante valiosas. Esta revisão identificou as mais utilizadas, demonstrando o estado da arte sobre técnicas de análise preditiva aplicadas à problemas relacionados às doenças respiratórias crônicas. Além disso, também foi possível observar que vários autores compararam o desempenho de múltiplas técnicas para alcançar o melhor resultado. Desta forma, destaca-se que a comparação de múltiplas técnicas é uma opção a ser considerada para a obtenção de melhores resultados na construção de modelos preditivos para essa área.

3.5 CONSIDERAÇÕES FINAIS

Foi realizada uma revisão sistemática, visando verificar na literatura técnicas de análise preditiva aplicadas à reabilitação pulmonar. Foram localizados 325 artigos nos motores de busca *Web of Science* e *Pubmed*. Desses, 20 foram selecionados para a fase final.

Os trabalhos selecionados para a fase final foram analisados para identificar o estado da arte em relação às técnicas de análise preditiva utilizadas na área de reabilitação pulmonar. Além disso, também foram verificados os algoritmos e técnicas mais utilizados, bem como verificadas as linguagens de programação e ferramentas ou *frameworks* para auxiliar na aplicação do modelo preditivo.

A técnica mais utilizada foi *Support Vector Machine*. Contudo, diversos autores compararam mais de uma técnica na busca de melhores resultados. O algoritmo que mais se destacou foi o *K-Nearest Neighbour Classifier*. Como forma de validação, grande parte dos autores utilizou sensibilidade e especificidade para validar o modelo. Também foi possível notar que foram utilizadas múltiplas formas de validação em alguns trabalhos para garantir um bom desempenho para o modelo. Em relação às linguagens de programação, a linguagem R demonstrou destaque.

Nesta revisão, não foram identificadas técnicas que foram aplicadas diretamente para a reabilitação pulmonar, confirmando a escassez de materiais relacionados à essa área. Os trabalhos identificados abordam, em geral, o diagnóstico, o grau de obstrução de vias aéreas, fenótipos e exacerbações de doenças respiratórias crônicas. Desta forma, foi identificada uma grande possibilidade de pesquisa na área de reabilitação pulmonar. Pode-se concluir que existe uma lacuna de pesquisa em relação à aplicação de técnicas de análise preditiva nessa área.

4 ANÁLISE PREDITIVA

A análise preditiva é uma tarefa de mineração de dados e aprendizado de máquina. Essa tarefa consiste na descoberta de relacionamentos entre exemplares de um conjunto de dados e rótulos a eles associados. Esse processo de descoberta de relacionamento é realizado por meio de um modelo de predição, que pode ser utilizado para predizer rótulos de exemplares desconhecidos após ajuste de seus parâmetros (SILVA; PERES; BOSCARIOLI, 2016).

O modelo preditivo é um modelo de classificação utilizado para prever o rótulo de determinadas instâncias de um conjunto de dados. As técnicas de classificação são bastante indicadas para conjuntos de dados com categorias nominais ou binárias. Existem diversas técnicas para a seleção do modelo preditivo. No entanto, algumas técnicas são mais indicadas para um tipo de problema específico (TAN; STEINBACH; KUMAR, 2009).

Alguns exemplos de áreas que utilizam essa técnica são análise de comportamento e expressão de emoções em redes sociais, biometria, tendências de ações no mercado financeiro, biologia, medicina, entre outras. Contudo, a aplicação de análise preditiva não se restringe apenas a essas áreas. Essa técnica é aplicável em um grande número de domínios (SILVA; PERES; BOSCARIOLI, 2016).

4.1 *MACHINE LEARNING*

Machine Learning (aprendizado de máquina) é a área da computação responsável por investigar formas de aprendizado dos computadores. A principal pesquisa da área é a pesquisa de métodos para que computadores reconheçam padrões a partir de dados existentes e tomem decisões inteligentes com base nos padrões descobertos (HAN; KAMBER; PEI, 2012). A tomada de decisões é realizada com a utilização de modelos matemáticos construídos a partir da teoria da estatística. Os modelos podem ser preditivos, fazendo previsões sobre o futuro, descritivos, extraindo conhecimento sobre os dados, ou ambos. Desta forma, a principal tarefa de *machine learning* é realizar inferências a partir de uma amostra (ALPAYDIN, 2010).

Segundo Alpaydin (2010), o aprendizado de máquina faz parte da área de inteligência artificial e é responsável por tratar problemas computacionais para os quais não existe um algoritmo, mas existem dados de exemplo. Para classificar um e-mail como *spam*, por exemplo, não existe um algoritmo, pois a classificação pode variar para cada indivíduo, não

existindo uma definição específica. A partir de um conjunto de e-mails que já foram classificados, é possível criar um modelo de aprendizado de máquina para aprender as características de um *spam* e classificar os próximos e-mails.

Machine Learning baseia-se em questões das áreas de Ciência da Computação e Estatística. A área da computação analisa questões como a construção de sistemas para solucionar problemas e quais são os problemas que podem ser tratados com computação. A área da estatística analisa questões como as inferências sobre dados e a confiabilidade com a qual essas inferências são realizadas. No entanto, são acrescentados alguns novos pontos, tais como fazer com que os computadores se programem automaticamente conforme a experiência adquirida e quais arquiteturas e algoritmos computacionais podem ser utilizados para isso (MITCHELL, 2006).

O desempenho do aprendizado de máquina é medido em relação à determinada atividade e o tipo de experiência. Pode-se afirmar que uma máquina aprende se seu desempenho ao realizar determinada atividade conforme a experiência adquirida melhorar de forma confiável. Portanto, o aprendizado de máquina dedica-se a responder questões sobre como é possível construir sistemas que melhoram automaticamente com a experiência e quais são os fatores que implicam nos processos de aprendizagem (MITCHELL, 2006).

4.1.1 Aprendizagem supervisionada

A aprendizagem supervisionada é praticamente um sinônimo de classificação. Nessa modalidade de aprendizado, é exigido um conjunto de dados de treinamento para o problema analisado. A aprendizagem é supervisionada devido a classificação de cada registro do conjunto de treinamento. Desta forma, o algoritmo aprende a partir do conjunto de treinamento fornecido. Um exemplo de aplicação dessa aprendizagem é a classificação de documentos de uma base de dados em artigos ou páginas da internet (HAN; KAMBER; PEI, 2012).

Além da classificação, a regressão também é um modelo de aprendizado supervisionado. Para esses modelos, existe uma entrada e uma saída e o modelo aprende o mapeamento da entrada para a saída conforme os dados de treinamento. O resultado é gerado conforme uma função discriminante (para os modelos de classificação) ou uma função de regressão (para os modelos de regressão), que separa as instâncias de classes diferentes. A partir de uma entrada, o modelo minimiza os erros de aproximação e, com isso, as estimativas

geradas são o mais próximas possível dos valores corretos do conjunto de treinamento (ALPAYDIN, 2010).

4.1.2 Aprendizagem não supervisionada

Diferentemente da aprendizagem supervisionada, na aprendizagem não supervisionada os dados não possuem atributos de classe. Esse modelo de aprendizagem é utilizado quando se deseja explorar os dados para encontrar padrões intrínsecos. A técnica de *clustering* (agrupamento) serve para identificar esses padrões. Ela consiste na divisão dos registros em grupos conforme a similaridade, chamados de *clusters* (LIU, 2011).

Na aprendizagem não supervisionada, o modelo possui somente os dados de entrada. A partir desses dados, é possível descobrir padrões frequentes. Por exemplo, se uma empresa possui dados de clientes e suas compras, é possível criar um modelo de aprendizagem não supervisionada para identificar padrões nos dados e criar grupos, segmentando os clientes conforme suas características de compras. Com isso, é possível criar estratégias de vendas conforme o comportamento dos clientes e até mesmo identificar clientes com comportamento diferente dos demais, indicando a existência de um novo nicho que pode ser explorado pela empresa (ALPAYDIN, 2010).

4.1.3 Aprendizagem semi-supervisionada

A aprendizagem semi-supervisionada mescla características dos modelos de aprendizagem supervisionado e não supervisionado. Os dados de entrada rotulados (aprendizagem supervisionada) são utilizados para aprender as classes das novas instâncias. Já os dados de entrada não rotulados (aprendizagem não supervisionada) são utilizados para refinar os limites entre as classes. Esse modelo de aprendizagem é indicado quando existe uma grande quantidade de dados não rotulados (HAN; KAMBER; PEI, 2012). Ele foi criado para reduzir o esforço manual de rotulagem, visto que, geralmente, a rotulagem é feita de forma manual (LIU, 2011).

4.1.4 Aprendizagem ativa

Visando otimizar a qualidade do modelo, a aprendizagem ativa permite que os usuários participem ativamente do processo de aprendizagem. Nessa abordagem, o usuário, um especialista por exemplo, pode ser solicitado a rotular uma instância ou um conjunto de

instâncias não rotuladas. Desta forma, o modelo usufrui de conhecimento humano para atingir melhores resultados (HAN; KAMBER; PEI, 2012).

Esse modelo é um tipo de aprendizado supervisionado. É indicado para situações em que existe uma grande quantidade de dados, mas com poucos rótulos. A aprendizagem é ativa na medida em que solicita ao usuário a rotulação de instâncias. Essa abordagem necessita de uma menor quantidade de instâncias classificadas do que o modelo de aprendizagem supervisionado clássico. O objetivo é atingir a maior precisão possível com o menor número possível de instâncias rotuladas (HAN; KAMBER; PEI, 2012). Assim como o modelo de aprendizagem semi-supervisionada, o modelo de aprendizagem ativa pode ser utilizado para reduzir o esforço manual de rotulagem (LIU, 2011).

4.2 TÉCNICAS DE ANÁLISE PREDITIVA

A grande quantidade de dados armazenada atualmente só se torna útil quando analisada e transformada em informações que podem ser utilizadas para gerar algum benefício, como, por exemplo, realizar previsões sobre um determinado processo. Existem padrões em dados armazenados. Sendo assim, supondo que o futuro próximo seja semelhante ao passado, os padrões coletados através dos dados armazenados podem ser utilizados para fazer previsões sobre o futuro (ALPAYDIN, 2010). Essa técnica é conhecida como **análise preditiva**.

Conforme Silva, Peres e Boscarioli (2016) destacam,

> A análise preditiva pode ser entendida como um processo que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributo de classe). Nesse contexto, um exemplar no conjunto de dados é um evento no domínio de análise (um prato servido no restaurante ou uma instância de um anúncio publicitário, por exemplo) e o rótulo pode ser apresentado de duas formas: (i) como identificação da classe à qual o evento está associado, dentro de um número finito de classes existentes no domínio de análise (pratos adequados para consumir com vinho branco ou pratos adequados para consumir com vinho tinto, por exemplo); (ii) como um número ao qual o evento está associado, dentre um conjunto contínuo de valores possíveis para essa associação (número de visualizações de um anúncio, por exemplo). A primeira forma de apresentação dos rótulos define uma situação para a análise preditiva do tipo classificação (ou predição categórica), e a segunda, uma situação para a análise preditiva do tipo regressão (ou predição numérica).

A análise preditiva é classificada como uma técnica de aprendizagem supervisionada e, portanto, depende de um conjunto de treinamento para o aprendizado. O processo de descoberta de relacionamento entre exemplares e rótulos é denominado modelo preditivo. A

partir do treinamento com o conjunto de teste, o modelo preditivo pode classificar exemplares que não fizeram parte do conjunto de dados usado para treinamento. Ou seja, classificar exemplares desconhecidos (SILVA; PERES; BOSCARIOLI, 2016). As cinco principais técnicas identificadas na revisão sistemática apresentada no Capítulo 3 serão apresentadas a seguir.

4.2.1 *Support Vector Machine*

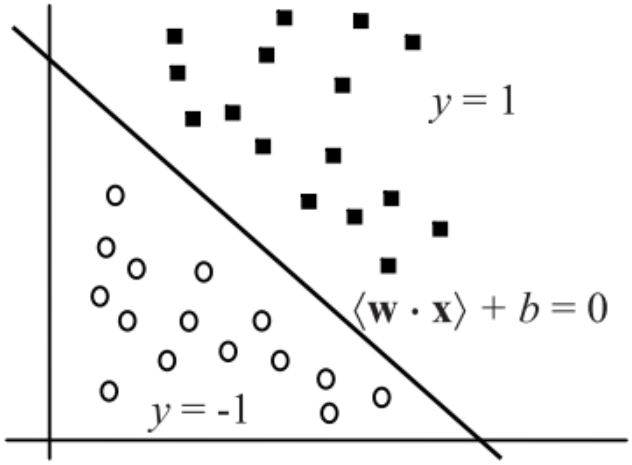
Support Vector Machine (SVM) é um dos algoritmos mais populares de *machine learning.* Sua precisão de classificação é maior que a maioria dos outros algoritmos em aplicações que envolvem dados de muitas dimensões. SVM é classificado como o algoritmo mais preciso para classificação de texto por diversos pesquisadores (LIU, 2011). Esse modelo é uma mescla de modelagem linear e aprendizado baseado em instâncias (WITTEN; FRANK, 2005).

Em geral, SVM constrói classificadores de duas classes (positiva e negativa) baseados em um sistema linear. A partir do conjunto de instâncias de dados, denominado vetor de entrada, o algoritmo identifica uma função linear para a construção de um modelo para classificação. A função linear é utilizada para gerar um limite entre as classes positiva e negativa, facilitando a classificação das instâncias. Desta forma, as instâncias que ficarem acima desse limite são atribuídas para a classe positiva e as instâncias que ficarem abaixo desse limite são atribuídas para a classe negativa (LIU, 2011).

O vetor de pesos é um parâmetro da função linear que é utilizado para formar o subconjunto de vetores de suporte, que são as instâncias que estão mais próximas do limite que separa as duas classes. Esses são os casos que são incertos e, por estarem próximos ao limite, permitem a extração de conhecimento. A partir da identificação dos vetores de suporte, é possível classificar as demais instâncias (ALPAYDIN, 2010).

A Figura 12 apresenta um gráfico de resultados gerados por um classificador SVM e a função linear identificada pelo modelo, onde **w** é o vetor de pesos, **x** é o vetor de entradas e **b** é o termo de tendência. A partir da função linear identificada, o modelo traça uma linha, que representa o limite entre as classes (positiva e negativa). No exemplo abaixo, a classe positiva é representada por quadrados, onde y é maior que zero. Já a classe negativa é representada por círculos, onde y é menor que zero.

Figura 12 – Gráfico representativo de resultados de classificação de um classificador SVM.



Fonte: Liu (2011)

4.2.2 *Decision Tree*

Decision Tree (DT) é uma estrutura de dados hierárquica que pode ser utilizada para classificação e regressão (ALPAYDIN, 2010). Em geral, os sistemas baseados em DTs têm boa precisão e a representação dos resultados em forma de árvore é intuitiva, facilitando o entendimento dos seres humanos sobre as regras geradas. Por isso, esses sistemas são bastante populares na área de *machine learning* (HAN; KAMBER; PEI, 2012).

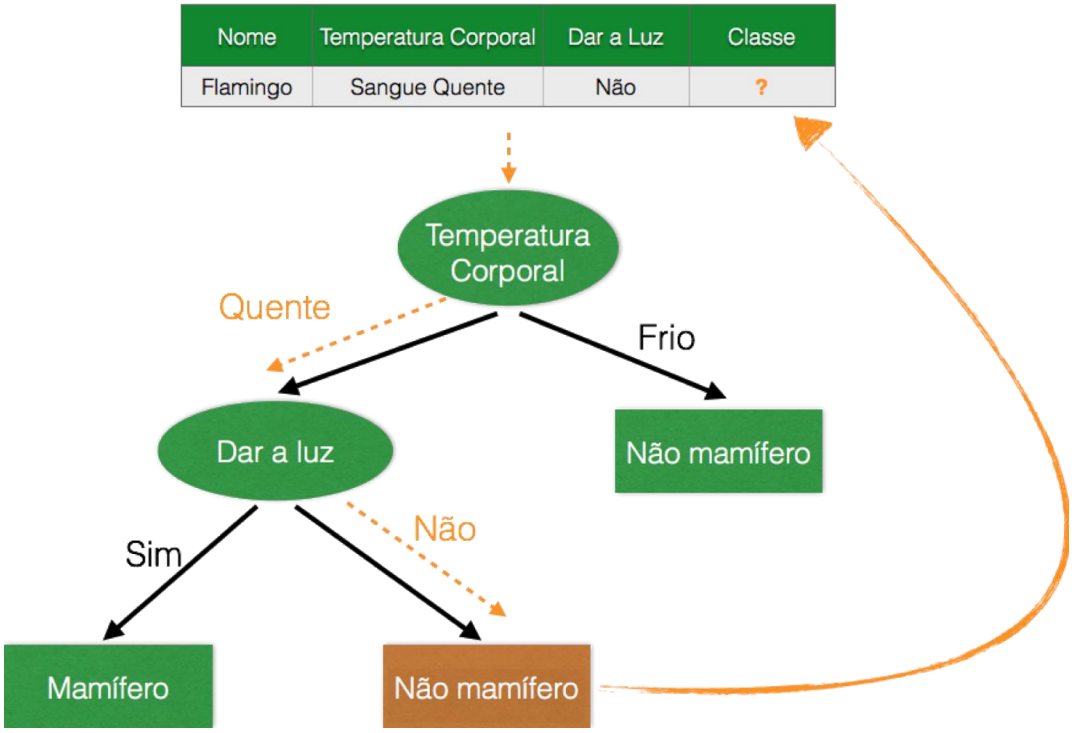
Esses sistemas geram um modelo preditivo com a estrutura de uma árvore e semelhante a um fluxograma. O nodo mais acima é denominado nodo raiz. Cada nodo interno representa o teste de um atributo das instâncias de dados. Cada ramificação gerada pelo nodo interno representa um resultado do teste. Cada nodo folha (terminal) representa um rótulo de uma classe que será atribuído para a instância analisada (HAN; KAMBER; PEI, 2012).

Mais de uma árvore pode ser gerada a partir de um mesmo conjunto de dados. Contudo, quanto menor for a árvore gerada, melhor o seu desempenho, pois a árvore é mais genérica. Uma árvore pequena também é facilmente entendida por usuários humanos. Isso é uma das vantagens da utilização de DT, pois, em alguns casos, o entendimento das regras por parte dos usuários é fundamental. Um exemplo disso é o diagnóstico de uma doença específica, onde os médicos necessitam saber quais os fatores que classificam se uma pessoa é portadora dessa doença (LIU, 2011).

A Figura 13 ilustra um exemplo de estrutura de árvore de decisão, onde o modelo tenta classificar se um determinado animal é mamífero. Os atributos selecionados para a geração do

modelo da árvore de decisão são “Temperatura corporal” (nodo raiz) e “Dar a luz”. Os rótulos (nodos terminais) que podem ser atribuídos para o atributo classe do modelo são “Mamífero” e “Não mamífero”. A partir de uma determinada instância, a árvore é percorrida até chegar em um nodo terminal, sendo esse o rótulo atribuído para a instância. No exemplo de árvore de decisão abaixo, o animal flamingo foi classificado como “Não mamífero”.

Figura 13 - Exemplo de estrutura de uma árvore de decisão.



Fonte: adaptado de Tan, Steinbach e Kumar (2009)

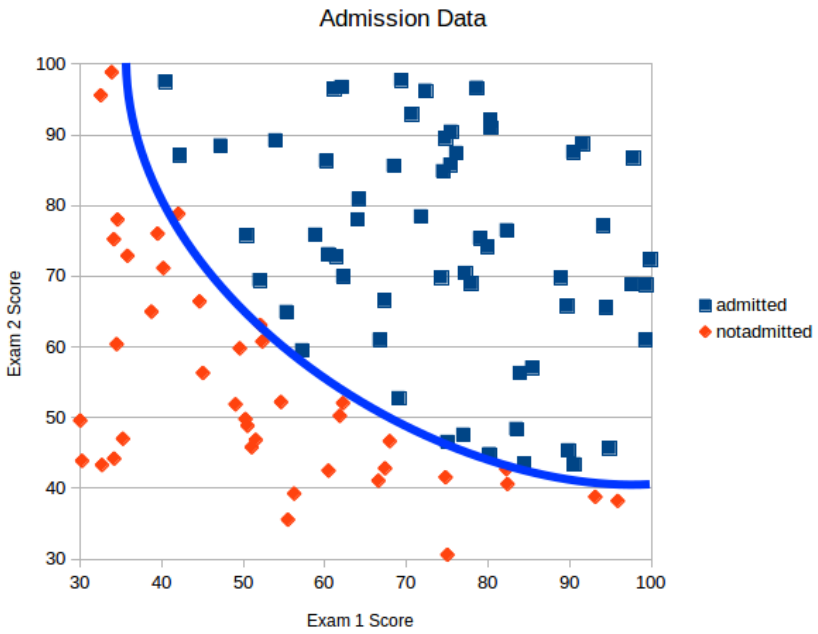
4.2.3 *Logistic Regression*

Logistic Regression (LR) é uma técnica baseada em regressão linear para classificação em domínios com atributos numéricos (WITTEN; FRANK, 2005). Segundo Pohar, Blas e Turk (2004, tradução nossa), “o objetivo da LR é encontrar o modelo mais adequado e mais econômico para descrever a relação entre o resultado (variável dependente ou resposta) e um conjunto de variáveis independentes (preditoras ou explicativas)”. Esse método é bastante robusto, flexível e fácil de usar. Os modelos de LR podem ser utilizados na saída de um modelo baseado em SVM para gerar estimativas de probabilidades (WITTEN; FRANK; 2005).

Os modelos de LR são indicados para a classificação linear e, portanto, estão relacionados a fronteiras para separar os rótulos que serão atribuídos para as instâncias

(POHAR; BLAS; TURK, 2004). Os modelos calculam estimativas de probabilidades. Desta forma, é possível fazer classificações precisas. A predição é realizada com a aproximação da variável desejada através da utilização de uma função linear. A partir disso, a instância é atribuída à classe a qual mais se aproxima (WITTEN; FRANK, 2005).

Figura 14 - Exemplo de classificação por regressão logística.



Fonte: Practicalai online (2017)

A Figura 14 ilustra um exemplo de um classificador baseado em LR. No exemplo abaixo, o modelo foi criado para calcular a probabilidade de um aluno ser aprovado em uma disciplina, classificando as instâncias (alunos) conforme o rótulo mais adequado (aprovado/*admitted* ou reprovado/*notadmitted*). Os eixos x e y representam as pontuações dos alunos nos exames 1 e 2, respectivamente. A linha traçada representa a função linear, que é o limite que divide as duas classes (aprovado e reprovado). Os alunos que estão acima da linha são classificados como aprovados (representados por quadrados), pois possuem nota acima da média da disciplina. Os demais alunos são classificados como reprovados (representados por losangos), pois possuem nota abaixo da média da disciplina.

4.2.4 *Random Forest*

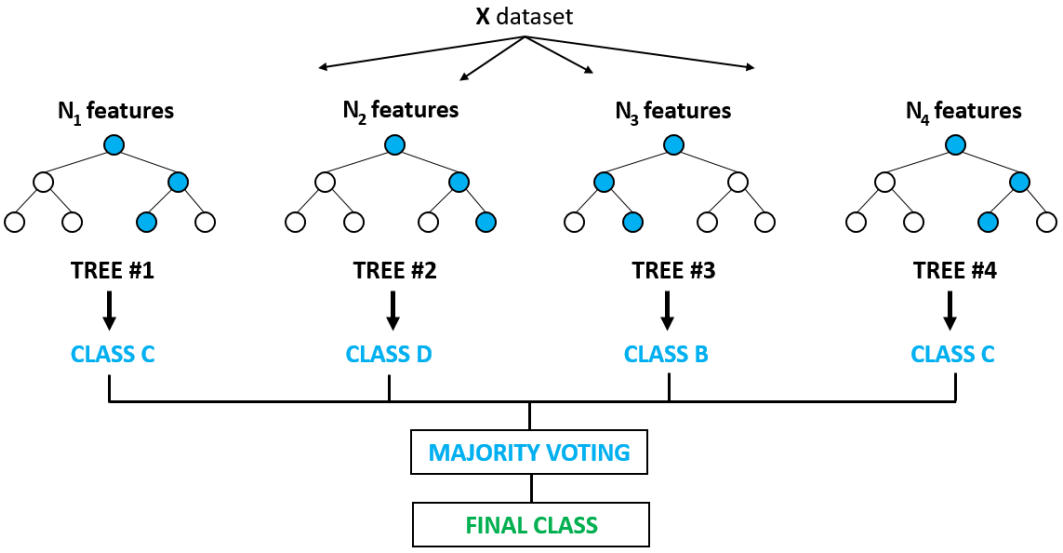
Random Forest (RF) é uma técnica de aprendizado de máquina que consiste em um conjunto de árvores de decisão. São geradas múltiplas árvores de decisão através de uma

seleção aleatória de atributos em cada nó. Por isso, a técnica é denominada “floresta”. Essa técnica é bastante robusta e eficiente em bancos de dados muito grandes. A precisão do conjunto é medida conforme a precisão dos classificadores individuais e a dependência entre eles. Para essa técnica, é extremamente indicado diminuir a correlação entre os classificadores individuais para que seja alcançado um bom desempenho (HAN; KAMBER; PEI, 2012).

Segundo Breiman (2001), RF é uma coleção de classificadores estruturados em árvore. Cada classificador individual ou árvore da floresta depende de um vetor aleatório, que é distribuído de forma equilibrada entre os classificadores. A partir de uma entrada, cada árvore de decisão classifica a instância individualmente. Ao final dessa etapa, as classificações são computadas e a classe mais votada é a classe atribuída para a instância.

Tan, Steinbach e Kumar (2009) dividem o processo de classificação através de um classificador RF em três passos. O primeiro passo consiste na geração de vetores randômicos após treinamento no conjunto de dados. O segundo passo equivale à criação de árvores de decisão aleatórias. Cada árvore é criada a partir de um vetor aleatório, que, por sua vez, é gerado a partir de uma distribuição de probabilidade fixa. Por fim, o passo três representa a classificação da instância baseada em uma estrutura de votação por maioria, onde a classe mais votada é atribuída para a instância.

Figura 15 - Exemplo de estrutura de um classificador RF.



Fonte: Holczer (2018)

A Figura 15 ilustra a estrutura de um classificador RF. Foram geradas 4 árvores de decisão diferentes com atributos selecionados aleatoriamente a partir de um mesmo conjunto de dados. Cada instância do conjunto de dados é analisada por todas as árvores de decisão da floresta e recebe um rótulo de cada árvore. O rótulo mais votado é atribuído para a instância. No exemplo abaixo, a classe C foi a mais votada, sendo assim atribuída para a instância analisada.

4.2.5 *Naïve Bayes*

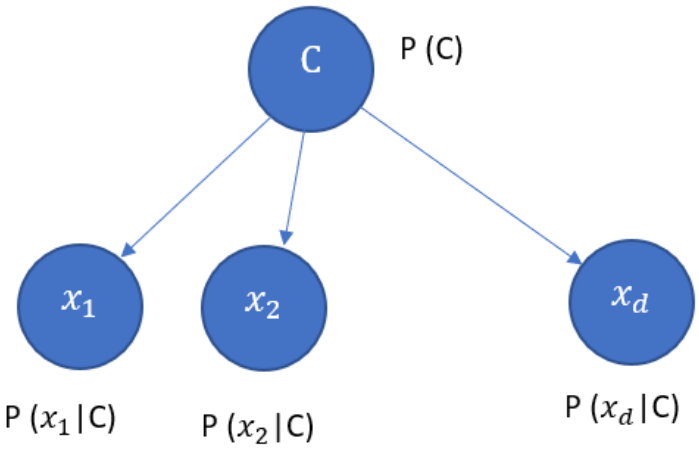
O classificador *Naïve Bayes* (NB) é um dos mais utilizados para classificação, pois tem um bom desempenho tanto com variáveis categóricas quanto com variáveis numéricas. Esse classificador foi criado com base no Teorema de Bayes (SILVA; PERES; BOSCARIOLI, 2016), que consiste em estimar a probabilidade de determinada hipótese ocorrer a partir da análise de uma evidência (HAN; KAMBER; PEI, 2012). Os classificadores bayesianos foram criados para solucionar problemas de classificação onde o rótulo de classe de um registro não pode ser previsto com certeza, mesmo que existam outros registros com os mesmos valores de atributos no conjunto de testes. Isso ocorre em problemas cujos dados possuem algum ruído ou existem fatores de confusão que afetam a classificação, mas não são utilizados na análise (TAN; STEINBACH; KUMAR, 2009).

O classificador NB calcula uma estimativa de probabilidade em vez de realizar previsões. Para cada classe, o classificador NB calcula a probabilidade de uma determinada instância pertencer a essa classe. Em geral, estimar a probabilidade é mais útil que apenas fazer uma previsão simples, pois é possível classificar as estimativas e minimizar seu custo. Uma das vantagens do classificador NB é que ele não fragmenta o conjunto de dados de treinamento e, portanto, consegue realizar estimativas mais confiáveis (WITTEN; FRANK, 2005).

Esse classificador é baseado em um modelo probabilístico. A classificação ocorre a partir de um cálculo para estimar a probabilidade, onde a classe com maior probabilidade é atribuída para a instância analisada (LIU, 2011). O classificador parte do princípio que os atributos são condicionalmente independentes, fazendo uma suposição ingênua. Por isso, leva esse nome. Apesar de simples, esses classificadores são bastante precisos e rápidos para grandes bancos de dados. A taxa de erro desses classificadores é pequena (MITRA; ACHARYA, 2003).

A Figura 16 apresenta o modelo de um classificador NB. Com base nos dados do conjunto de treinamento, são realizados os cálculos probabilísticos a fim de gerar o modelo para a classificação de novas instâncias. Primeiramente, é calculada a probabilidade de ocorrência de cada classe C (rótulo). Após isso, cada instância é analisada e, para os valores de cada atributo descritivo referente à instância, é calculada a probabilidade desses valores resultarem na classificação da instância com a classe C. Esse cálculo é realizado para todos os atributos descritivos em relação aos rótulos de cada instância do conjunto de testes. Cada atributo é analisado individualmente conforme o princípio de que os atributos são independentes. Por fim, as probabilidades são multiplicadas e a classe com maior probabilidade é atribuída para a instância.

Figura 16 - Exemplo de estrutura de um classificador NB.



Fonte: adaptado de Alpaydin (2010)

4.3 LINGUAGENS DE PROGRAMAÇÃO

Algumas linguagens de programação podem ser utilizadas para a aplicação de análise preditiva sobre bases de dados, bem como para gerar visualizações sobre esses dados. Em geral, essas linguagens possuem pacotes que podem ser instalados para facilitar a aplicação das técnicas de análise preditiva. R[1] e Python[2] são exemplos de linguagens que podem ser utilizadas para a análise de dados. Conforme a revisão sistemática apresentada no Capítulo 3, essas são as linguagens de programação mais utilizadas para a análise preditiva na área de doenças respiratórias crônicas.

[1] Página para download do R: https://www.r-project.org/
[2] Página para download do Pyhton: https://www.python.org/

R é uma linguagem e um ambiente de programação para computação estatística e geração de gráficos. R disponibiliza uma grande variedade de técnicas para análise estatística, tais como as modelagens linear e não linear, classificação e agrupamento. Essa linguagem também fornece várias técnicas para facilitar a plotagem de gráficos, sendo essa uma de suas principais vantagens. R está disponível como *software* livre e é altamente extensível (THE R FOUNDATION, 2018).

O pacote "randomForest"[3] é um exemplo de pacote disponível para R (LEIDY *et al.*, 2016). Esse pacote auxilia na aplicação de técnicas de classificação e regressão com base em classificadores RF. Para regressão logística, pode-se utilizar o pacote "glmnet"[4] (ZHONG *et al.*, 2017). Esse pacote pode ser utilizado para outros modelos lineares além da regressão logística, como, por exemplo, a regressão de Poisson e o modelo de Cox. Outro pacote que pode ser utilizado é o "pamr"[5] (GUAN *et al.*, 2016), que possui funções para classificação de amostras em *microarrays*.

O RStudio[6] é uma ferramenta que pode ser utilizada para auxiliar na programação com R. A principal vantagem dessa ferramenta é a interface gráfica, o que facilita o desenvolvimento de aplicações (SILVA; PERES; BOSCARIOLI, 2016). O RStudio também possui uma edição disponível como *software* livre (RSTUDIO, 2018).

Python é uma linguagem de programação interpretada, interativa e orientada a objetos. É uma linguagem extensível e pode ser utilizada como linguagem de extensão para sistemas que necessitam de uma interface programável. Python é portátil, permitindo que seja executado em diversos sistemas operacionais. Essa linguagem de programação também é disponibilizada como *software* livre (PYTHON SOFTWARE FOUNDATION, 2018).

Python é uma linguagem de alto nível que pode ser aplicada à diversos problemas. A instalação padrão disponibiliza uma grande biblioteca que atende à diversas áreas, tais como processamento de *strings*, protocolos de internet, engenharia de software e interfaces do sistema operacional. Além disso, existe uma ampla variedade de extensões de terceiros disponível para instalação (PYTHON SOFTWARE FOUNDATION, 2018).

---

[3] Página do pacote: https://cran.r-project.org/web/packages/randomForest/index.html
[4] Página do pacote: https://cran.r-project.org/web/packages/glmnet/index.html
[5] Página do pacote: https://cran.r-project.org/web/packages/pamr/index.html
[6] Página para download do RStudio: https://www.rstudio.com/

O pacote “Scikit-Learn”[7] é um pacote Pyhton para *machine learning*. Esse pacote disponibiliza técnicas de aprendizado supervisionado e não supervisionado para classificação, regressão e agrupamento. Esse pacote foi desenvolvido para não especialistas em *machine learning*, dando ênfase para a facilidade no uso e para o desempenho (PEDREGOSA *et al.*, 2011). Estudos recentes (SPATHIS; VLAMOS, 2017) (SWAMINATHAN *et al.*, 2017) mencionaram a utilização desse pacote a fim de auxiliar na construção de um modelo preditivo para o diagnóstico de doenças respiratórias crônicas.

4.4 CONSIDERAÇÕES FINAIS

A análise preditiva é uma tarefa de mineração de dados e aprendizado de máquina que consiste na descoberta de relacionamento entre instâncias de um conjunto de dados e os rótulos a elas associados. Nessa tarefa, é utilizado um modelo preditivo para classificar instâncias de um conjunto de dados por meio da predição dos rótulos. Esse modelo de classificação é mais indicado para conjuntos de dados com categorias nominais ou binárias.

Existem várias técnicas para aprendizado de máquina. Elas são classificadas conforme o tipo de aprendizado. Na aprendizagem supervisionada, o aprendizado ocorre a partir de um conjunto de treinamento. Na aprendizagem não supervisionada, o aprendizado ocorre através da identificação de padrões internos de um conjunto de dados. A aprendizagem semi-supervisionada mescla características das aprendizagens supervisionada e não supervisionada.

O processo de predição é realizado por meio de modelos preditivos. Esses, por sua vez, são criados com a utilização de técnicas de análise preditiva. *Support Vector Machine* é uma técnica utilizada para construir classificadores de duas classes baseados em um sistema linear. *Decision Tree* gera modelos com estrutura de árvores, facilitando a compreensão dos seres humanos sobre as regras geradas. *Logistic Regression* é uma técnica baseada em regressão linear e tem como objetivo encontrar o modelo mais adequado e econômico. *Random Forest* gera um conjunto de árvores de decisão. *Naïve Bayes* é baseado em um modelo probabilístico, que calcula a probabilidade de uma determinada instância pertencer a determinada classe.

Algumas linguagens de programação podem ser utilizadas para a análise preditiva em conjuntos de dados e para gerar gráficos. R e Python são exemplos de linguagens que podem

[7] Página do pacote: http://scikit-learn.org/stable/index.html

auxiliar nesses processos. Essas linguagens também são extensíveis, possibilitando a instalação de pacotes que facilitam a aplicação das técnicas.

5 APLICAÇÃO DE ANÁLISE PREDITIVA NA REABILITAÇÃO PULMONAR

Com base na revisão sistemática e no estudo aprofundado sobre as técnicas de *machine learning* apresentados nos capítulos anteriores, foram selecionadas as técnicas SVM, DT e RF para comparação de desempenho. As técnicas SVM e DT foram selecionadas por serem as duas mais frequentes identificadas na revisão sistemática. A técnica RF foi a quarta mais frequente identificada na revisão e foi selecionada pela grande similaridade com a técnica DT, facilitando sua aplicação. O abandono foi escolhido para a análise preditiva por se tratar de um grande problema atual do projeto de extensão.

Após a escolha das técnicas, iniciou-se a etapa de pré-processamento. Com os dados ajustados, foram aplicadas as etapas de modelagem, validação e comparação de desempenho entre as técnicas. Por fim, foi desenvolvida uma ferramenta para aplicação das técnicas e visualização de informações sobre os dados analisados. Todas as etapas serão detalhadas a seguir.

5.1 PRÉ-PROCESSAMENTO

Tendo em vista que os dados do projeto de extensão são armazenados em uma planilha Excel, foi necessária uma etapa de ajuste dos dados na fase de pré-processamento para torná-los aptos para a análise. Esse tipo de armazenamento ocasiona diversos problemas na coleta dos dados, tais como diferentes formas de escrita de uma mesma palavra, valores não informados por falta de obrigatoriedade de preenchimento, problemas de digitação, entre outros. Além disso, alguns atributos não são coletados para todos os pacientes pelo fato de determinado paciente não ter realizado algum exame ou por ter abandonado o tratamento ainda na fase de coleta de dados iniciais.

5.1.1 Seleção de atributos

Dentre os 356 atributos coletados atualmente, foram selecionados 10 para a análise. Assim como realizado por Swaminathan *et al.* (2017), este trabalho utilizou preditores já conhecidos na área estudada para a geração dos modelos preditivos. Esse processo de seleção foi realizado com apoio da professora coordenadora do projeto de extensão (especialista na área), que indicou os atributos mais relevantes. Para essa seleção, foram avaliados somente os atributos que têm sua coleta realizada logo no início do tratamento, visto que esses demonstram, em geral, o estado inicial dos pacientes ao ingressar no tratamento. Após a

seleção dos atributos, foi gerado um novo conjunto de dados contendo somente os atributos selecionados. A nomenclatura dos atributos desse conjunto foi ajustada para remover os caracteres especiais. O dicionário dos dados analisados está disponível no Apêndice A deste trabalho.

Os atributos numéricos selecionados foram “Espirometria VEF1%” (percentual do volume expiratório do primeiro segundo medido na espirometria simples), “Gênero” (indicador do gênero do paciente, onde “1” indica gênero “Masculino” e “2” indica gênero “Feminino”), “Idade” (idade do paciente), “AC_Quantas internações no último ano?” (quantidade de internações do paciente no último ano), “Comorbidades HAS” (indicador de presença de Hipertensão Arterial Sistêmica, onde “1” indica “Sim” e “2” indica “Não”), “Comorbidades Diabetes” (indicador de presença de Diabetes, onde “1” indica “Sim” e “2” indica “Não”), “Comorbidades Cardiopatias” (indicador de presença de Doenças Cardíacas, onde “1” indica “Sim” e “2” indica “Não”) e “Escala de Dispnéia MRC_Antes” (indicador de gravidade da falta de ar coletado através do formulário Escala de Dispneia MRC). Os atributos descritivos selecionados foram “Doença” (doença portada pelo paciente) e “Comorbidades Outras” (indicador de presença de outras comorbidades, onde o valor “2” indica “Não” e os valores preenchidos indicam a presença das doenças informadas).

Por falta de um atributo indicador de abandono do tratamento, fez-se necessária a criação do atributo “Abandonou” para essa finalidade. Esse atributo foi preenchido com base nos atributos “QQVSG Sint.Depois” (indicador do resultado no formulário baseado no Questionário do Hospital Saint George do paciente após a conclusão do tratamento) e “TCSatO2i Depois” (indicador da saturação de oxigênio do paciente após a conclusão do tratamento). Esses atributos são coletados obrigatoriamente na finalização do tratamento. A ausência de valores nesses atributos para determinado paciente indica abandono do tratamento. Do contrário, o paciente finalizou o tratamento.

5.1.2 Limpeza de dados

Os dados armazenados contêm diversos valores ausentes. Em geral, os valores não foram armazenados devido ao abandono do tratamento antes mesmo da coleta dessas informações. Para a aplicação das técnicas SVM e RF, o conjunto de dados não pode conter valores ausentes em nenhum atributo. Desta forma, foi necessário um processo de limpeza de dados, visando remover os valores ausentes.

Primeiramente, foi avaliado o atributo “Doença”, que não possuía valor em sete registros. Com o apoio da professora coordenadora do projeto de extensão, esses valores foram preenchidos manualmente na planilha de dados. Com base nos valores de espirometria dos pacientes, dois registros foram preenchidos com o valor “DPOC” e três registros foram preenchidos com o valor “SEM DPOC”. Os dois registros restantes não possuíam valores de espirometria. Desta forma, foram preenchidos com o valor “NÃO INFORMADO” para evitar desconsiderá-los.

Após isso, foram analisados os atributos “Comorbidades HAS”, “Comorbidades Diabetes”, “Comorbidades Cardiopatias” e “Comorbidades Outras”, que possuíam 8, 9, 9 e 103 valores ausentes, respectivamente. Esses registros também foram preenchidos com apoio da professora coordenadora do projeto de extensão. Os valores ausentes foram preenchidos com o valor “NÃO”, indicando que não há a presença de tal comorbidade.

Por fim, foram analisados os demais atributos numéricos que possuíam valores ausentes. Com base em uma das possíveis abordagens para completar valores ausentes destacada por Faceli *et al.* (2011), os valores ausentes desses atributos foram preenchidos com um valor numérico padrão diferente para cada atributo, indicando que o atributo possuía um valor desconhecido. Os 80 valores ausentes do atributo “Espirometria VEF1%” foram preenchidos com o valor “150”. O atributo “Idade” foi ajustado com o valor “120”. Esse atributo possuía 3 valores ausentes. Para esse atributo, também foi realizado um experimento de aplicação das técnicas de análise preditiva utilizando a média, para não afetar a distribuição desse atributo no conjunto de dados. O desempenho, porém, foi inferior ao do experimento utilizando o valor padrão. O atributo “AC_Quantas internações no último ano?” possuía 89 valores ausentes e foi ajustado com o valor “-1”. Finalmente, o atributo “Escala de Dispnéia MRC_Antes” foi ajustado com o valor “5”. Esse atributo possuía 215 registros sem valores. Para o ajuste de cada atributo, foi selecionado um valor fora da faixa de valores possíveis para tal atributo ou um valor com pouca probabilidade de ocorrência, evitando que seja interpretado como um valor coletado por meios normais durante o tratamento.

5.1.3 Transformação de dados

Alguns dos atributos utilizados para a análise são preenchidos com valores numéricos para indicar um valor descritivo, como, por exemplo, o atributo “Gênero” cujos valores

possíveis são “1” e “2” para indicar os gêneros “Masculino” e “Feminino”, respectivamente. Atributos com essa situação foram ajustados para facilitar a interpretação dos resultados.

Os valores do atributo “Gênero” foram ajustados para “MASCULINO” e “FEMININO” conforme a indicação de cada valor numérico. Os valores dos atributos “Comorbidades HAS”, “Comorbidades Diabetes”, “Comorbidades Cardiopatias” e “Comorbidades Outras” foram ajustados para “SIM” e “NÃO” conforme a indicação dos valores numéricos ou o preenchimento do campo.

Também na etapa de transformação de dados, foi realizado um trabalho em conjunto com a professora coordenadora do projeto de extensão para padronização dos valores do atributo “Doença”. No conjunto de dados original, existem 50 valores diferentes para indicar a doença portada pelos pacientes. Após esse trabalho de padronização, o número de valores foi reduzido para 27.

5.2 MODELAGEM

Finalizada a etapa de pré-processamento, foi gerado um novo conjunto de dados ajustado para modelagem e comparação de desempenho entre as técnicas de *machine learning.* Foi gerado um modelo preditivo para cada técnica, considerando todos os atributos do conjunto de dados. Tendo em vista que o atributo “Escala de Dispnéia MRC_Antes” possuía muitos valores ausentes que foram preenchidos com um valor padrão, também foi gerado um segundo modelo preditivo para cada técnica, desconsiderando esse atributo. O objetivo dos modelos é prever a classe do atributo “Abandonou”, sendo a classe positiva “NÃO” e a classe negativa “SIM”. A linguagem de programação R foi escolhida para a modelagem, visto que é uma linguagem bastante utilizada para essa área como identificado na revisão sistemática apresentada no Capítulo 3.

Para fins de comparação, foi utilizado o mesmo conjunto de dados para a modelagem de todas as técnicas. Cada modelo preditivo foi treinado e testado com as técnicas *10-fold cross validation* (validação cruzada), que é uma técnica frequentemente utilizada em trabalhos na área como apresentado na revisão sistemática descrita no Capítulo 3, e *holdout*. Foram gerados dois subconjuntos de dados balanceados com 70% e 30% dos registros para treinamento e teste, respectivamente. Essas proporções de particionamento também foram utilizadas por Finkelstein e Jeong (2017). O particionamento foi realizado com o pacote R caret. Nos modelos testados com validação cruzada, a técnica *10-fold cross validation* foi

aplicada sobre o conjunto de treinamento. Após isso, foi realizada a validação sobre os dados de teste. Os modelos validados pela técnica *holdout* foram treinados apenas uma única vez, utilizando o conjunto de treinamento. Para validação final dessa técnica, foi utilizado o conjunto de teste.

5.2.1 *Support Vector Machine*

Os modelos foram gerados com a utilização dos pacotes R caret e e1071. O primeiro modelo foi gerado considerando todos os atributos do conjunto de dados gerado após o pré-processamento. O segundo modelo foi gerado desconsiderando o atributo "Escala de Dispnéia MRC_Antes". Os dois primeiros modelos foram treinados com a utilização da técnica *10-fold cross validation* sobre os dados de treinamento. Também foram gerados mais dois modelos treinados apenas uma vez sobre os dados de treinamento, totalizando quatro modelos.

O *kernel* é um dos parâmetros que pode ser ajustado nos modelos SVM e consiste em uma fórmula matemática utilizada para identificar a melhor separação entre as classes. O *kernel* linear pode ser utilizado para classificar dados linearmente separáveis. Também existem outros tipos para separação de dados não separáveis linearmente, tais como *polynomial, Gaussian radial basis* e *sigmoid.* Não existem regras para seleção de um *kernel* que resultará em um modelo mais preciso. Em geral, o *kernel* não gera grande diferença no desempenho do modelo, pois essa técnica identifica uma solução global (HAN; KAMBER, 2006). Neste trabalho, foi utilizado o *kernel* linear em todos os modelos para possibilitar a compatibilidade entre os pacotes.

O pacote R caret foi utilizado para aplicar a técnica *10-fold cross validation.* Esse pacote também auxiliou na identificação do melhor valor para o parâmetro "cost", utilizado para penalizar o modelo no caso de classificações incorretas. Para isso, foram testados todos os valores no intervalo de 0,02 a 0,5, variando em 0,02. O melhor valor para o parâmetro para cada modelo foi escolhido com base na melhor acurácia gerada. O valor 0,02 foi o que gerou melhor acurácia para o modelo que considerou todos os atributos do conjunto de dados. Já para o modelo que desconsiderou o atributo "Escala de Dispnéia MRC_Antes", o melhor valor para o parâmetro foi 0,06. O pacote R e1071 foi utilizado para aplicar a técnica *holdout* nos dois últimos modelos. Nesses, foi utilizado o mesmo valor do parâmetro "cost" dos respectivos modelos treinados com validação cruzada, visando replicar o mesmo cenário em todas as comparações.

5.2.2 *Decision Tree*

Os modelos foram gerados com a utilização dos pacotes R caret e rpart. O primeiro modelo foi gerado considerando todos os atributos do conjunto de dados gerado após o pré-processamento. O segundo modelo foi gerado sem a utilização do atributo "Escala de Dispnéia MRC_Antes". Os dois primeiros modelos foram treinados com a utilização da técnica *10-fold cross validation* sobre os dados de treinamento. Foram gerados, ainda, mais dois modelos treinados apenas uma vez sobre os dados de treinamento, gerando um total de quatro modelos.

Para a aplicação da técnica *10-fold cross validation,* foi utilizado o pacote R caret. Esse pacote também auxiliou na busca dos melhores valores para determinados parâmetros dos modelos. Primeiramente, foi identificado o melhor valor para o parâmetro "split", que define como os dados serão organizados, visando o ganho de informação. Para esse parâmetro, foram testados os valores "Information" (entropia) e "Gini" (índice de Gini). Outro parâmetro ajustado foi o "minsplit", que define a quantidade mínima de exemplares em um nó para que ele tenha subárvores. Foram testados todos os valores inteiros entre 1 e 20. Foi utilizada a acurácia como base para seleção do melhor valor para ambos os parâmetros. Tanto no modelo gerado com os dados completos quanto no modelo que desconsiderou o atributo "Escala de Dispnéia MRC_Antes", o melhor valor para o parâmetro "split" foi "Information". Desta forma, ambos os modelos utilizaram a entropia como forma de ganho de informação. O valor 14 foi utilizado no parâmetro "minsplit" de ambos os modelos. O pacote R rpart foi utilizado para aplicar a técnica *holdout* nos dois últimos modelos. Os melhores valores para os parâmetros, que foram identificados no treinamento dos modelos com validação cruzada, também foram utilizados nos respectivos modelos treinados uma única vez.

5.2.3 *Random Forest*

Os modelos foram gerados com a utilização dos pacotes R caret e randomForest. O primeiro modelo foi gerado considerando todos os atributos do conjunto de dados gerado após o pré-processamento. O segundo modelo foi gerado desconsiderando o atributo "Escala de Dispnéia MRC_Antes". Os dois primeiros modelos foram treinados com a utilização da técnica *10-fold cross validation* sobre os dados de treinamento. Também foram gerados mais

dois modelos treinados apenas uma vez sobre os dados de treinamento, totalizando quatro modelos.

O pacote R caret foi utilizado para aplicar a técnica *10-fold cross validation* no treinamento dos modelos. O pacote também foi utilizado para identificar o melhor valor para alguns parâmetros com base na acurácia gerada. Para o parâmetro “mtry”, que define o número de preditores selecionados aleatoriamente utilizados para a geração de cada nó, foram testados todos os valores inteiros entre 1 e 10. Os valores 250, 300, 350, 400, 450, 500, 550, 600, 800, 1000 e 2000 foram testados para identificar o número de árvores geradas, definido pelo parâmetro “ntree”. Os valores 9 e 350 foram identificados como melhores valores para os parâmetros “mtry” e “ntree”, respectivamente, no modelo que utilizou todos os atributos. No modelo que desconsiderou o atributo “Escala de Dispnéia MRC_Antes”, foram identificados os valores 2 e 450 como melhores valores para os parâmetros “mtry” e “ntree”, respectivamente. Foi utilizado o pacote R randomForest para aplicar a técnica *holdout* nos dois últimos modelos. Para esses modelos, foram utilizados os mesmos valores dos parâmetros dos respectivos modelos treinados com validação cruzada.

5.3 VALIDAÇÃO

A última etapa antes do início do desenvolvimento da ferramenta foi a de validação dos modelos preditivos gerados. Conforme mencionado anteriormente, as técnicas *10-fold cross validation* e *holdout* foram selecionadas para validação dos modelos. A técnica de validação cruzada foi selecionada por ser uma técnica bastante utilizada em trabalhos na área como demonstrado na revisão sistemática apresentada no Capítulo 3. Também com base nos resultados da revisão sistemática, foram selecionadas as métricas de validação de desempenho: sensibilidade, especificidade e área sob a curva *Receiver Operating Characteristic* (ROC). Juntamente à essas métricas, foram utilizadas acurácia e precisão. A curva ROC foi utilizada para visualização gráfica do desempenho dos modelos devido à utilização da área sob a curva ROC como medida de desempenho.

As métricas selecionadas foram utilizadas para medir o desempenho de cada modelo, permitindo uma comparação detalhada entre os modelos. Inicialmente, foi realizada uma comparação entre os quatro modelos gerados para cada técnica. Após isso, foi realizada uma comparação entre os modelos treinados com *10-fold cross validation* de todas as técnicas. Desta forma, a comparação final avaliou o desempenho de seis modelos.

5.3.1 Técnicas de validação de desempenho

A técnica *holdout* consiste na divisão do conjunto de dados em duas partes com as proporções de p e $1 – p$. A primeira parte é utilizada para treinamento do modelo e a segunda para teste. Normalmente, é utilizado $p = 2/3$. Um dos problemas apresentados por essa técnica é não possibilitar a análise da variação de desempenho entre conjuntos de treinamentos com registros diferentes. O desempenho de um modelo preditivo pode variar dependendo de quais registros são utilizados para formar o conjunto de treinamento, visto que esse é gerado com base em uma proporção sobre o conjunto de dados originais (FACELI *et al.*, 2011).

Na técnica *k-fold cross validation,* os dados são divididos em k subconjuntos aleatórios com tamanho aproximado. Após isso, são realizados treinamento e teste k vezes. A cada vez, um subconjunto é utilizado para teste e os demais para treinamento. Desta forma, os subconjuntos são utilizados o mesmo número de vezes para treinamento e uma vez para teste. Um dos objetivos da divisão é manter a distribuição de classes do conjunto de dados original em cada subconjunto. Em geral, *10-fold cross validation* é recomendada para estimar a acurácia de um modelo devido à baixa variação (HAN; KAMBER, 2006). A divisão em 10 subconjuntos tornou-se padrão após vários testes em diversos conjuntos de dados com diferentes técnicas de *machine learning* mostrarem que esse número é bastante adequado para obter a melhor estimativa de erro. Esses testes também mostraram que a divisão melhora consideravelmente os resultados. Por mais que o número padrão de divisões seja 10, as divisões em 5 ou 20 subconjuntos também proporcionam resultados melhores (WITTEN; FRANK; HALL, 2011).

5.3.2 Métricas de validação de desempenho

A **matriz de confusão** apresenta a quantidade de acertos e erros de um determinado modelo preditivo. É um instrumento bastante útil para a medida de desempenho. É formada por colunas que representam os verdadeiros positivos, os falsos positivos, os falsos negativos e os verdadeiros negativos. Utilizando como exemplo um problema de duas classes, positiva e negativa, os **verdadeiros positivos** são os registros da classe positiva que foram corretamente classificados pelo classificador. Os **falsos positivos** são os registros da classe negativa que foram incorretamente classificados com a classe positiva, bem como os **falsos negativos** são os registros da classe positiva que foram incorretamente classificados com a classe negativa. Finalmente, os **verdadeiros negativos** são os registros da classe negativa que foram

corretamente classificados (HAN; KAMBER, 2006). A Figura 17 apresenta a estrutura de uma matriz de confusão.

Figura 17 - Estrutura da matriz de confusão para um problema de duas classes.

	Valor Real		
Valor Previsto	Classe	Positiva	Negativa
	Positiva	Verdadeiros Positivos	Falsos Positivos
	Negativa	Falsos Negativos	Verdadeiros Negativos

Fonte: adaptado de Han e Kamber (2006)

A partir da matriz de confusão, são calculadas algumas métricas de validação de desempenho. A **sensibilidade** corresponde à taxa de acertos na classe positiva. Ela também é chamada de taxa de verdadeiros positivos. Seu valor é calculado dividindo o número de verdadeiros positivos pela soma entre o número de verdadeiros positivos e o número de falsos negativos. A **especificidade** é a taxa de acertos na classe negativa. Também é conhecida por taxa de verdadeiros negativos. Seu valor é calculado dividindo o número de verdadeiros negativos pela soma entre o número de verdadeiros negativos e o número de falsos positivos. A **acurácia** é a taxa de acerto total de um classificador. É calculada pela soma dos valores da diagonal principal da matriz, verdadeiros positivos e verdadeiros negativos, dividida pela soma dos valores de todos os elementos da matriz. A **precisão** é a taxa de acerto de registros classificados com a classe positiva dentre todos os registros que foram classificados com a classe positiva. Ou seja, seu valor é calculado dividindo o valor de verdadeiros positivos pela soma entre o número de verdadeiros positivos e o número de falsos positivos (FACELI *et al.*, 2011).

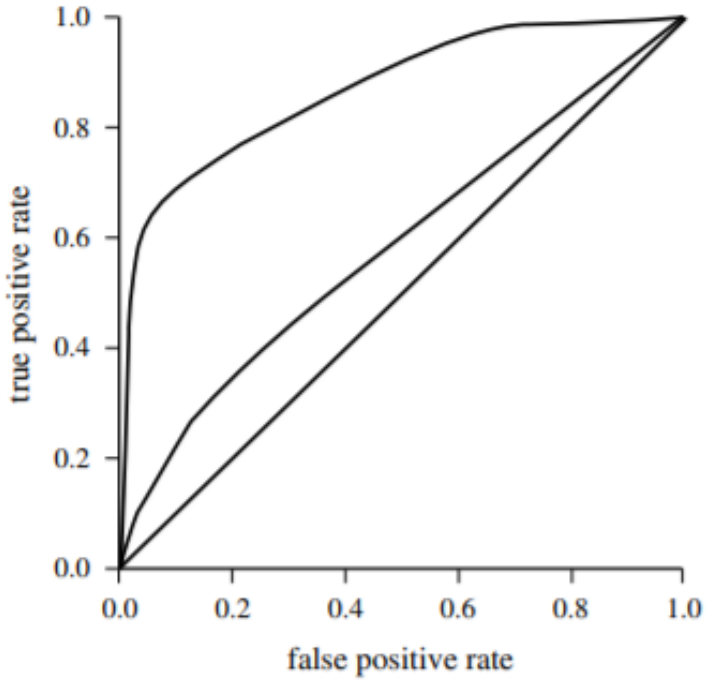
5.3.3 Análise ROC

Gráficos são instrumentos valiosos para análises. A **curva ROC** demonstra o desempenho de um classificador, apresentando a taxa de verdadeiros positivos no eixo vertical e a taxa de verdadeiros negativos no eixo horizontal. Quanto mais próximo ao canto superior esquerdo do gráfico, melhor é o desempenho do classificador. A sigla significa característica de operação do receptor, termo utilizado na área de detecção de sinais para

apresentar a relação entre a taxa de acerto e a taxa de alarme falso em um canal com ruído (WITTEN; FRANK; HALL, 2011).

A curva ROC apresenta a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos, que é a proporção de registros da classe negativa que foram classificados incorretamente sobre o número total de registros da classe negativa (1 – especificidade). Tendo em vista essa relação, um aumento na taxa de verdadeiros positivos implica no aumento da taxa de falsos positivos. Além disso, essas representações gráficas são excelentes para comparar o desempenho de dois modelos de classificação (HAN; KAMBER, 2006).

Figura 18 - Exemplo de Curvas ROC de dois modelos de classificação.



Fonte: Han e Kamber (2006)

Para gerar uma curva ROC, os registros do conjunto de testes devem ser ordenados de forma decrescente de acordo com a probabilidade de classificação para a classe prevista. A curva deve iniciar no canto inferior esquerdo, onde ambas as taxas estão zeradas. Para cada registro, é validada sua classe prevista. No caso de acerto, a curva deve ser movida para cima e deve ser desenhado um ponto, aumentando a taxa de verdadeiros positivos. Do contrário, a curva deve ser movida para a direita e deve ser desenhado um ponto, aumentando a taxa de falsos positivos. Esse processo deve ser repetido para todos os registros do conjunto de teste (HAN; KAMBER, 2006). A Figura 18 ilustra as curvas ROC de dois modelos de

classificação. A figura também apresenta uma linha diagonal, onde, para cada verdadeiro positivo, tem-se a mesma probabilidade de encontrar um falso positivo.

Para facilitar a comparação de desempenho de múltiplos modelos quando existirem interseções entre as curvas, é comum a utilização de uma medida única extraída de cada curva ROC. Essa medida é denominada **área sob a curva ROC (AUC)**. Essa medida produz valores entre 0 e 1, indicando a área abaixo da curva ROC. Quanto maior o valor, melhor o desempenho do modelo. Na comparação entre dois classificadores, o que possuir maior área sob a curva ROC é o que possui melhor desempenho (FACELI *et al.*, 2011).

5.3.4 Comparação de desempenho entre os modelos preditivos

Com base nas métricas e técnicas de validação selecionadas, foi realizada uma comparação de desempenho entre os modelos preditivos. Inicialmente, realizou-se uma comparação entre os modelos preditivos de cada técnica de *machine learning* utilizada. Após isso, realizou-se uma comparação de desempenho geral entre os modelos treinados com *10-fold cross validation*, visando identificar o melhor modelo preditivo. Todas as medidas apresentadas a seguir foram geradas após a validação de cada modelo sobre um subconjunto de teste que representa 30% do conjunto de dados original. Nenhum registro do conjunto de teste foi utilizado no treinamento dos modelos. O pacote R caret foi utilizado para o cálculo das métricas acurácia, sensibilidade, especificidade e precisão. O pacote R ROCR foi utilizado para o cálculo da AUC e para a geração das curvas ROC.

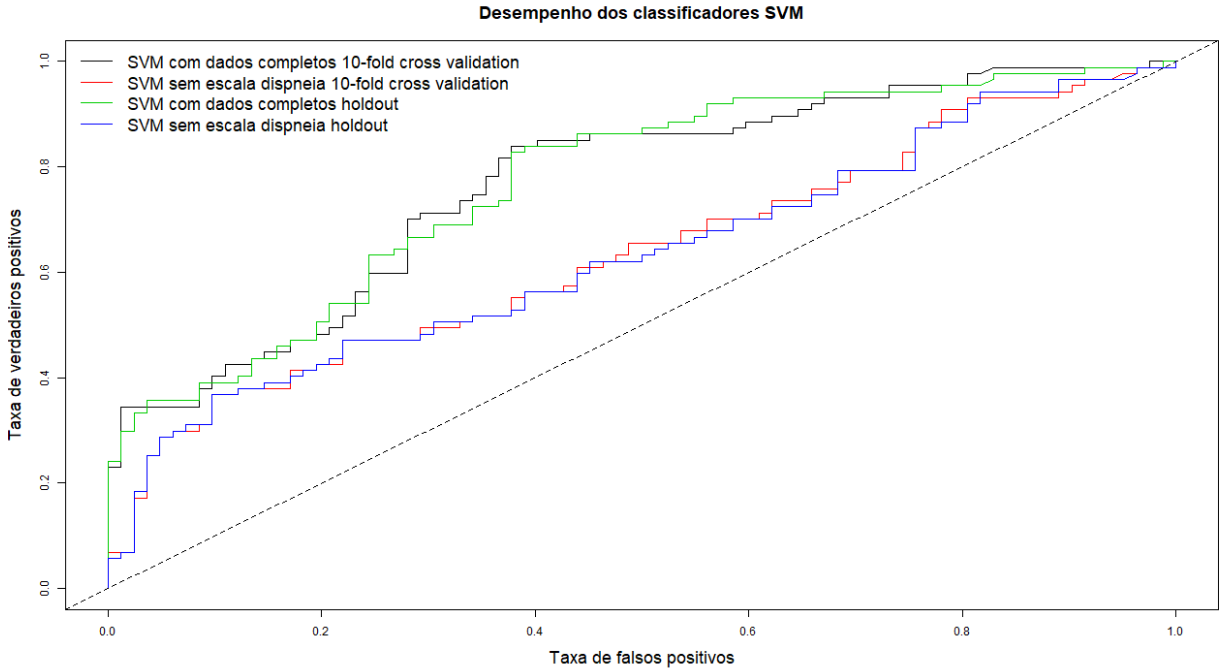
O Quadro 10 apresenta as medidas de desempenho dos modelos preditivos gerados com a aplicação da técnica SVM. Pode-se observar que o modelo gerado considerando todos os atributos do conjunto de dados e treinado uma única vez obteve a melhor acurácia, bem como a melhor especificidade, a melhor precisão e a melhor AUC. Ambos os modelos que desconsideraram o atributo “Escala de Dispnéia MRC_Antes” apresentaram sensibilidade excelente, ultrapassando 0,9. A especificidade apresentada por esses modelos, no entanto, foi muito baixa, gerando um baixo desempenho geral. Neste trabalho, a especificidade representa maior importância, pois indica a quantidade de acertos de pacientes que efetivamente abandonaram o tratamento. A Figura 19 apresenta as curvas ROC dos modelos analisados.

Quadro 10 - Comparação de desempenho dos modelos preditivos SVM.

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	AUC
SVM com dados completos 10-fold cross validation	0.6746	0.7561	0.5977	0.6392	0.7684
SVM sem escala dispneia 10-fold cross validation	0.6095	0.9024	0.3333	0.5606	0.6362
SVM com dados completos holdout	0.6864	0.6951	0.6782	0.6706	0.7696
SVM sem escala dispneia holdout	0.6036	0.9024	0.3218	0.5564	0.6337

Fonte: elaborado pelo autor

Figura 19 - Curvas ROC dos modelos preditivos SVM.



Fonte: elaborado pelo autor

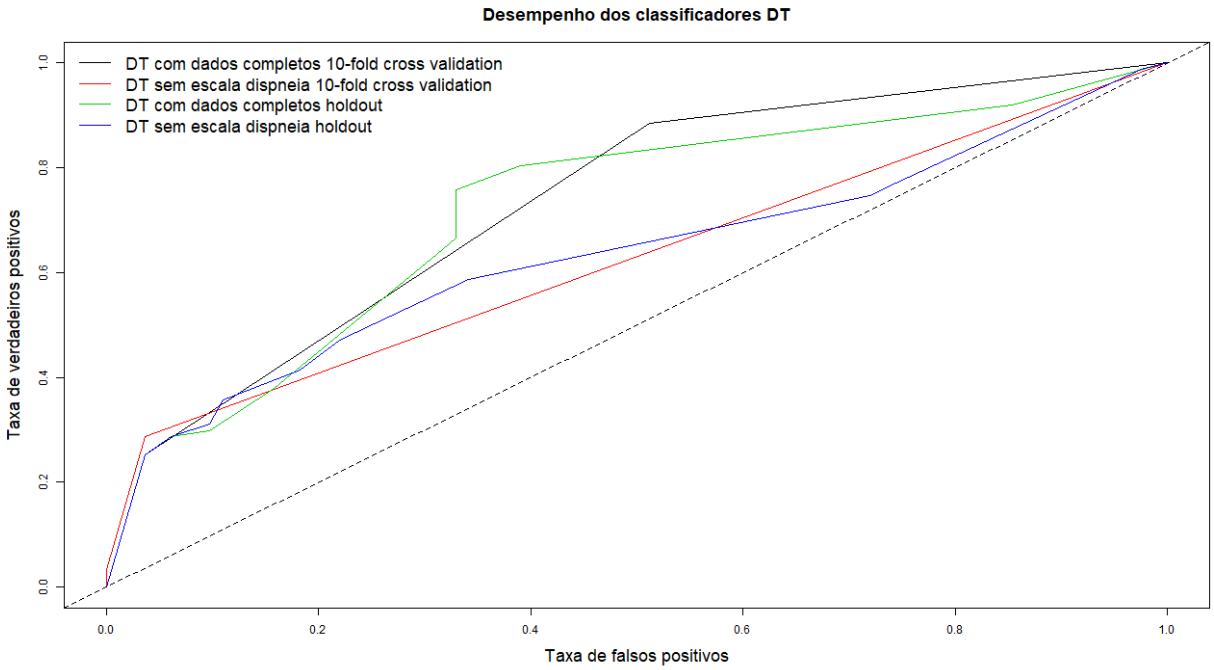
As medidas de desempenho dos modelos preditivos gerados com a técnica DT são apresentadas no Quadro 11. Os modelos que utilizaram todos os atributos do conjunto de dados obtiveram o maior desempenho em relação à acurácia. No entanto, o modelo treinado com validação cruzada obteve melhor especificidade, precisão e AUC. O modelo treinado com validação cruzada que desconsiderou o atributo "Escala de Dispnéia MRC_Antes" atingiu disparadamente a melhor sensibilidade, chegando em quase 100% de acerto. Contudo, obteve uma especificidade extremamente baixa, impactando no desempenho geral do modelo. Esse modelo também apresentou as medidas mais baixas nas demais métricas. A Figura 20 apresenta as curvas ROC dos modelos.

Quadro 11 - Comparação de desempenho dos modelos preditivos DT.

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	AUC
DT com dados completos 10-fold cross validation	0.6923	0.4878	0.8851	0.8	0.7350
DT sem escala dispneia 10-fold cross validation	0.6154	0.9634	0.2874	0.5603	0.6260
DT com dados completos holdout	0.6923	0.6707	0.7126	0.6875	0.7193
DT sem escala dispneia holdout	0.6213	0.6585	0.5862	0.6	0.6336

Fonte: elaborado pelo autor

Figura 20 - Curvas ROC dos modelos preditivos DT.



Fonte: elaborado pelo autor

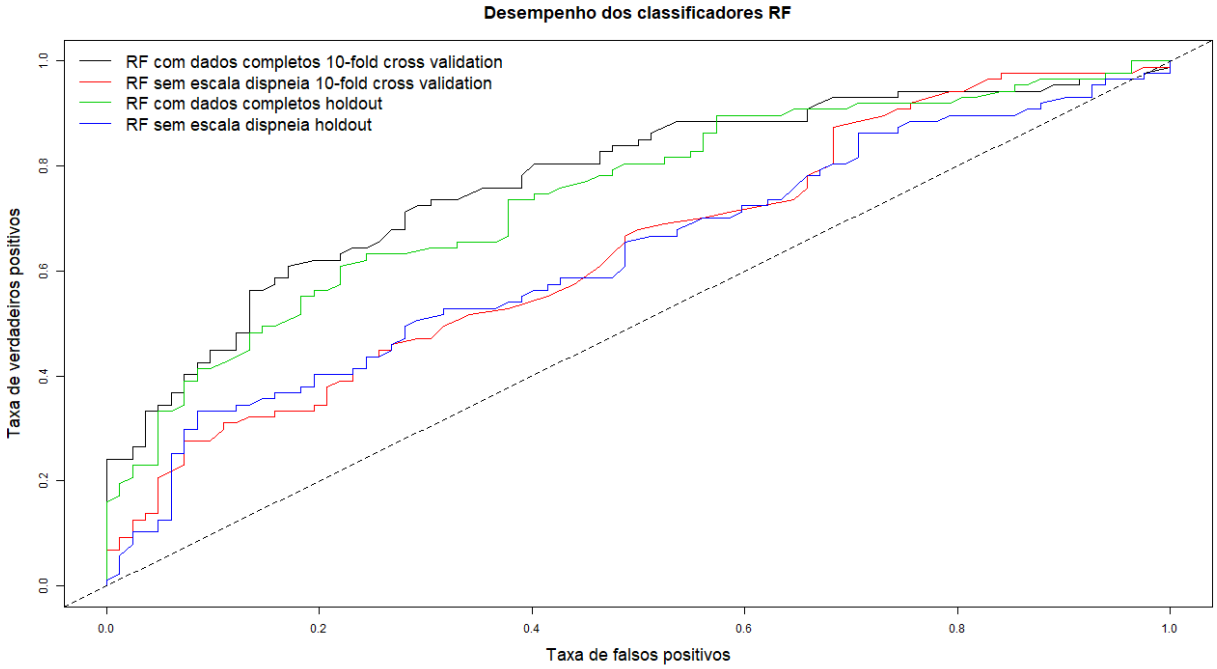
O Quadro 12 apresenta a comparação de desempenho entre os modelos preditivos gerados com a aplicação da técnica RF. O modelo gerado com a utilização de todos os atributos do conjunto de dados e treinado com a técnica *10-fold cross validation* obteve a melhor acurácia. Esse modelo também apresentou melhor precisão, melhor AUC e segunda melhor sensibilidade. O respectivo modelo treinado com a técnica *holdout* apresentou resultados aproximados, tendo melhor especificidade. Desta forma, pode-se avaliar que os modelos que consideraram o atributo “Escala de Dispnéia MRC_Antes” apresentaram melhor desempenho em relação aos respectivos modelos que desconsideraram esse atributo. A Figura 21 apresenta as curvas ROC dos modelos analisados.

Quadro 12 - Comparação de desempenho dos modelos preditivos RF.

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	AUC
RF com dados completos 10-fold cross validation	0.6982	0.7683	0.6322	0.6632	0.7735
RF sem escala dispneia 10-fold cross validation	0.5740	0.8415	0.3218	0.5391	0.6370
RF com dados completos holdout	0.6627	0.6707	0.6552	0.6471	0.7429
RF sem escala dispneia holdout	0.5917	0.7195	0.4713	0.5619	0.6272

Fonte: elaborado pelo autor

Figura 21 - Curvas ROC dos modelos preditivos RF.



Fonte: elaborado pelo autor

Após a comparação de desempenho entre os modelos preditivos de cada técnica, foi realizada uma comparação geral. O Quadro 13 apresenta a comparação de desempenho entre os modelos treinados com validação cruzada de todas as técnicas de análise preditiva utilizadas. O modelo gerado com a técnica RF que utilizou os dados completos obteve melhor acurácia, embora a diferença para o respectivo modelo gerado com a técnica DT seja pequena. Pode-se observar que, em relação à acurácia, os modelos que utilizaram os dados completos atingiram resultados melhores. A melhor sensibilidade foi atingida pelo modelo gerado com a técnica DT que desconsiderou o atributo “Escala de Dispnéia MRC_Antes”. A melhor especificidade foi alcançada pelo modelo gerado com a técnica DT que utilizou os dados completos. De modo geral, os modelos que utilizaram os dados completos obtiveram melhor

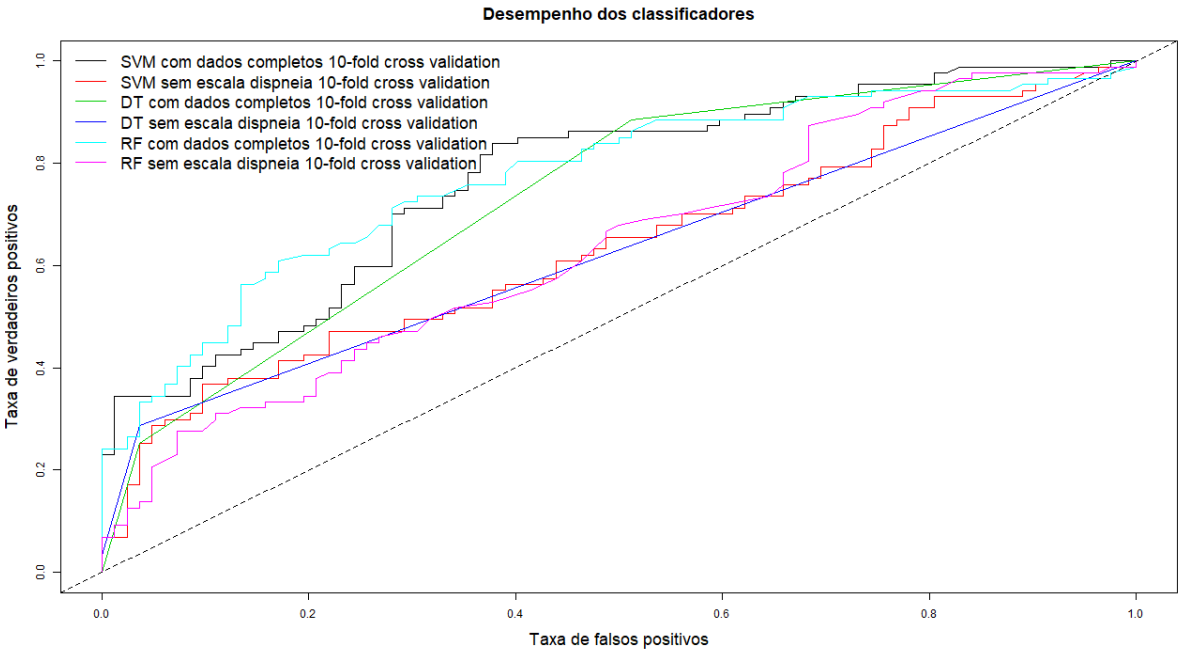
especificidade. Em contrapartida, os modelos que não utilizaram os dados completos obtiveram melhor sensibilidade, mas obtiveram especificidade extremamente baixa. Em relação à precisão, os modelos que utilizaram os dados completos alcançaram desempenho melhor. Por fim, os modelos que utilizaram os dados completos tiveram maior área atingida na curva ROC, gerando melhor AUC. A Figura 22 ilustra as curvas ROC dos modelos analisados.

Quadro 13 - Comparação de desempenho geral dos modelos preditivos.

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	AUC
SVM com dados completos 10-fold cross validation	0.6746	0.7561	0.5977	0.6392	0.7684
SVM sem escala dispneia 10-fold cross validation	0.6095	0.9024	0.3333	0.5606	0.6362
DT com dados completos 10-fold cross validation	0.6923	0.4878	0.8851	0.8	0.7350
DT sem escala dispneia 10-fold cross validation	0.6154	0.9634	0.2874	0.5603	0.6260
RF com dados completos 10-fold cross validation	0.6982	0.7683	0.6322	0.6632	0.7735
RF sem escala dispneia 10-fold cross validation	0.5740	0.8415	0.3218	0.5391	0.6370

Fonte: elaborado pelo autor

Figura 22 - Curvas ROC dos modelos preditivos.



Fonte: elaborado pelo autor

5.4 CONSIDERAÇÕES FINAIS

O processo de aplicação das técnicas de *machine learning* sobre os dados do projeto de extensão foi dividido em três etapas. Primeiramente, foi realizada a etapa de pré-processamento dos dados. Após isso, foi realizada a geração dos modelos preditivos na etapa de modelagem. Por fim, foi realizada a etapa de validação, na qual o desempenho dos modelos preditivos foi comparado através de métricas de desempenho.

Na etapa de pré-processamento, foram selecionados os atributos que foram utilizados para a análise preditiva. Também foi realizada uma etapa de limpeza de valores ausentes. Para isso, foram adotadas algumas estratégias para o preenchimento dos valores ausentes. Os valores ausentes do atributo “Doença” foram preenchidos com base nos valores de espirometria do paciente. Os valores ausentes dos atributos referentes às comorbidades foram preenchidos com “NÃO”, indicando a não existência da comorbidade. Esse processo foi realizado com auxílio da professora coordenadora do projeto de extensão. Os demais atributos, “Espirometria VEF1%”, “Idade”, “AC_Quantas internações no último ano?” e “Escala de Dispnéia MRC_Antes”, foram preenchidos com valores numéricos padrões fora do intervalo de valores possíveis ou com pouca probabilidade de ocorrência.

Na etapa de modelagem, foram criados os modelos preditivos para cada técnica com o auxílio da linguagem de programação R. Para cada técnica, foi gerado um modelo considerando todos os atributos do conjunto de dados e outro modelo desconsiderando o atributo “Escala de Dispnéia MRC_Antes”. Para treinamento e teste de cada modelo, foram utilizadas as técnicas *10-fold cross validation* e *holdout.* Com isso, quatro modelos preditivos foram criados para cada técnica.

A etapa de validação consistiu na coleta das métricas de desempenho acurácia, sensibilidade, especificidade, precisão e AUC de cada modelo. Além disso, foram geradas curvas ROC para comparação de desempenho. Inicialmente, foram comparados os quatro modelos de cada técnica. Após essa comparação, foi realizada uma comparação de desempenho geral entre os modelos treinados com *10-fold cross validation* de cada técnica. Nessa comparação geral, foram analisados seis modelos preditivos. Em relação à acurácia, o modelo gerado considerando todos os atributos e treinado uma única vez obteve o melhor desempenho entre os modelos da técnica SVM. Os dois modelos que foram gerados considerando todos os atributos obtiveram o mesmo desempenho entre os modelos da técnica DT, atingindo a melhor acurácia. Entre os modelos da técnica RF, o modelo gerado

considerando todos os atributos e treinado com a técnica *10-fold cross validation* obteve o melhor desempenho. Esse modelo também obteve a maior acurácia na comparação de desempenho geral.

6 DESENVOLVIMENTO DA FERRAMENTA

Finalizada a comparação entre os modelos, foi iniciada a última etapa, que consistiu no desenvolvimento de uma ferramenta[8] para aplicação das técnicas de *machine learning* e visualização de informações sobre os dados do projeto de extensão. A ferramenta foi desenvolvida com a utilização da linguagem de programação R. Os pacotes shiny e shinydashboard foram utilizados para o desenvolvimento de painéis interativos em R. O pacote DT foi utilizado para a geração de tabelas. Para a geração os gráficos, foram utilizados os pacotes googleVis e plotly. O gráfico referente à árvore de decisão foi gerado com o pacote rpart.plot.

Foi criada uma tela de *login* para garantir a segurança dos dados do projeto de extensão disponibilizados na ferramenta. O painel de visualização foi dividido em três guias principais, sendo elas "Novos Pacientes", "Estatísticas" e "Base de Dados". A divisão de guias da ferramenta é apresentada na Figura 23.

Figura 23 - Divisão de guias do painel de visualização da ferramenta.

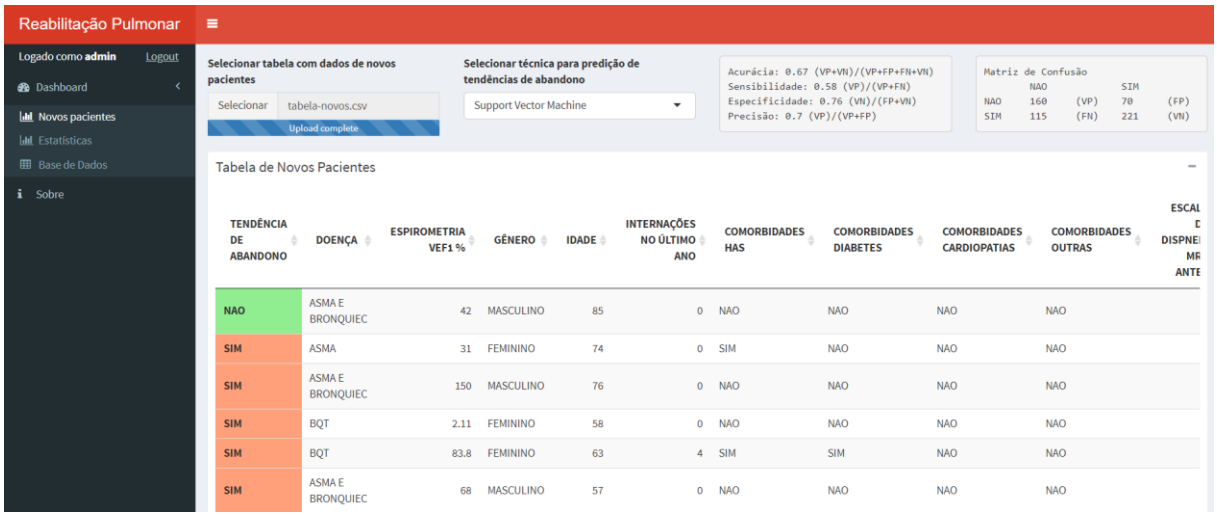
Fonte: elaborado pelo autor

A guia "Novos Pacientes" possibilita a importação de uma tabela de pacientes para prever a tendência de abandono dos pacientes que estão ingressando no tratamento. A tabela deve possuir os mesmos atributos utilizados para a modelagem apresentados na Subseção

[8] Página da ferramenta: http://ceted.feevale.br:3838/rp/

5.1.1. Optou-se por disponibilizar as três técnicas de *machine learning* na ferramenta para que o usuário possa definir qual técnica deseja utilizar para a análise. Para auxiliar na escolha da técnica, são apresentadas visualmente no canto superior direito as métricas de desempenho de cada técnica, assim como a matriz de confusão gerada pelo modelo. Tanto as métricas como a matriz de confusão são geradas com base no teste de cada modelo sobre os dados de treinamento. Após a importação da tabela com novos pacientes, o resultado da análise é apresentado em forma de tabela, onde a coluna "Tendência de Abandono" corresponde ao atributo-alvo dos modelos preditivos. A tabela é atualizada quando a técnica de *machine learning* é alterada. A Figura 24 ilustra a guia "Novos Pacientes" após a importação de uma tabela com novos pacientes.

Figura 24 - Guia de novos pacientes após importação de tabela com novos pacientes.



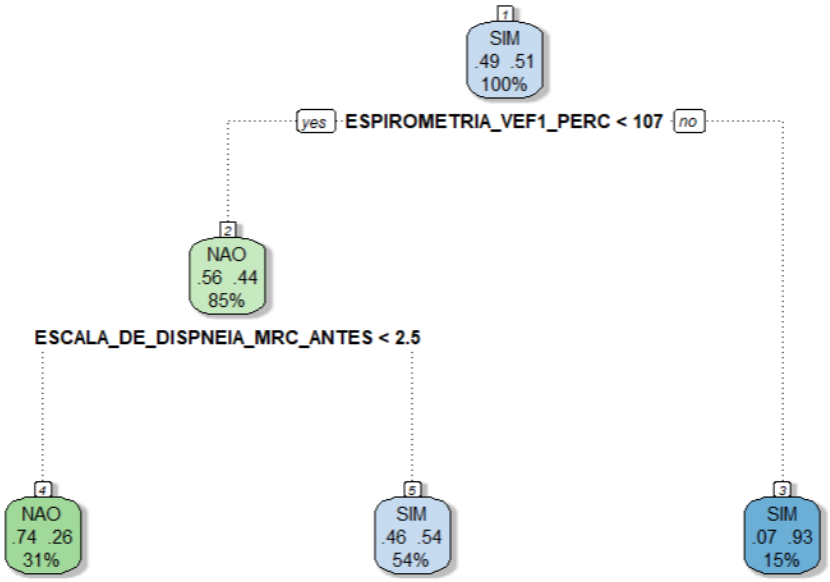
Fonte: elaborado pelo autor

Os modelos preditivos são gerados automaticamente por meio de uma tarefa agendada no servidor devido ao tempo médio de processamento de 10 minutos. Os modelos das três técnicas são gerados com a utilização de todos os atributos do conjunto de dados e treinados com a técnica *10-fold cross validation.* Os parâmetros são ajustados na geração de cada modelo. Esse processo faz com que os modelos estejam sempre ajustados aos dados atualizados. A cada geração, os modelos são salvos em um arquivo com extensão "RData", que é importado na inicialização da ferramenta.

Nessa guia, também são exibidas visualizações sobre a tabela importada, apresentando a distribuição dos novos pacientes, tais como gráfico de barras com a contagem de pacientes por gênero, gráfico de caixa *(boxplot)* com a distribuição de idade por gênero e gráfico de

barras com a contagem de pacientes por doença. Também são apresentadas as visualizações referentes aos modelos preditivos DT e RF.

Figura 25 - Árvore de decisão gerada.

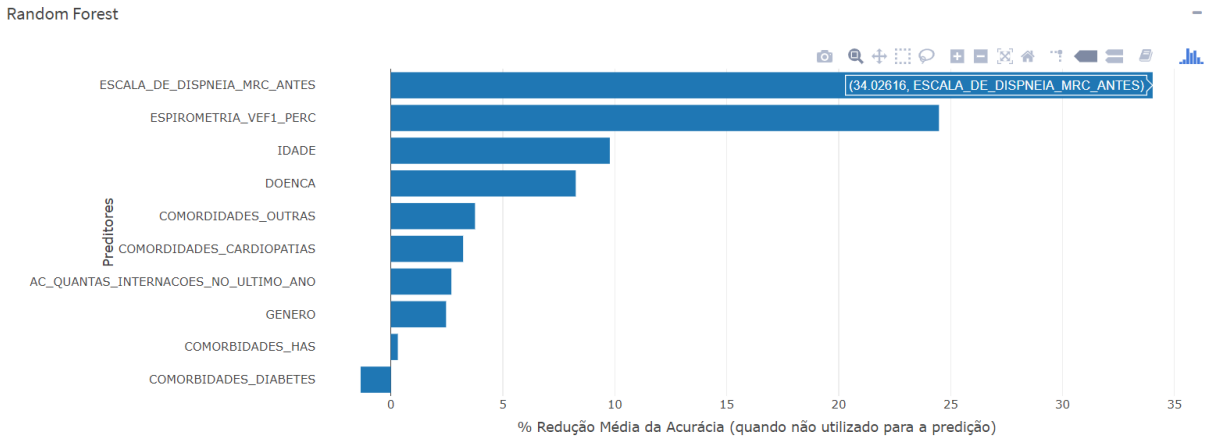


Fonte: elaborado pelo autor

Em relação ao modelo DT, a árvore gerada pelo modelo é apresentada na Figura 25. Cada nó final especifica a categoria que o modelo atribuirá para os registros que entrarem naquele nó. Os pacientes com espirometria VEF1 % menor que 107 e escala de dispneia menor que 2,5 serão classificados como “NÃO” pelo modelo, indicando que não possuem tendências de abandono. Já os pacientes que possuem espirometria VEF1 % menor que 107 e escala de dispneia maior ou igual a 2,5 ou espirometria VEF1 % maior que 107 serão classificados pelo modelo como “SIM”, indicando que possuem tendências de abandono. A árvore também apresenta a proporção de elementos em cada nó analisado. No nó raiz, no qual é validado o atributo “ESPIROMETRIA_VEF1_PERC”, foram analisados 100% dos registros do conjunto de treinamento. No segundo nó, no qual é validado o atributo “ESCALA_DE_DISPNEIA_MRC_ANTES”, 85% dos registros foram analisados e assim por diante. O modelo também apresenta, em cada nó não terminal, a classe majoritária do atributo-alvo, “ABANDONOU”, e a proporção de registros de cada classe desse atributo no conjunto de treinamento. Por exemplo, no nó raiz, 51% dos pacientes analisados abandonaram o tratamento e 49% não abandonaram.

Em relação ao modelo RF, é apresentado um gráfico (Figura 26) com os preditores (atributos) mais importantes, indicando o percentual de redução média da acurácia quando o preditor é excluído do modelo. O modelo RF calcula a medida de importância de cada atributo do conjunto de dados. Todos os gráficos da guia “Novos Pacientes” são demonstrados no Apêndice B deste trabalho.

Figura 26 - Preditores mais importantes para o modelo RF.



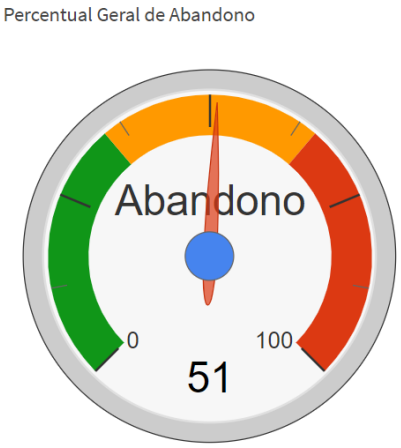
Fonte: elaborado pelo autor

A guia “Estatísticas” apresenta visualizações interativas sobre todos os registros do conjunto de dados. Nessas visualizações, não são considerados os novos pacientes, levando em consideração que ainda não é sabido se esses abandonaram ou não o tratamento. As visualizações disponíveis nessa guia envolvem todos os atributos do conjunto de dados, buscando gerar estatísticas relevantes para o projeto de extensão. Alguns dos gráficos dessa guia foram sugeridos pela professora coordenadora do projeto. Os gráficos exibidos nessa guia abordam o percentual geral de abandono do projeto, a proporção de doenças, a distribuição de espirometria VEF1 % por idade, a distribuição de idade por gênero, a distribuição da escala de dispneia por espirometria VEF1 % e gênero, quantidade de abandonos geral e por gênero, distribuição de escala de dispneia, contagem de pacientes por comorbidade e sem comorbidade, quantidade de abandonos por escala de dispneia e média de internações no último ano por doença. Todos os gráficos dessa guia são demonstrados no Apêndice C deste trabalho.

A primeira visualização, conforme ilustrado pela Figura 27, apresenta o percentual geral de abandono. Atualmente, 51% dos pacientes abandonaram o tratamento, o que representa um número de 291 pacientes. Os motivos para o abandono variam entre falecimento, internação, mudança de endereço e outros fatores desconhecidos. O período de

abandono também varia de acordo com cada paciente. Alguns abandonos ocorrem logo no início do tratamento, antes mesmo da realização de exames iniciais para identificação do perfil de saúde do paciente.

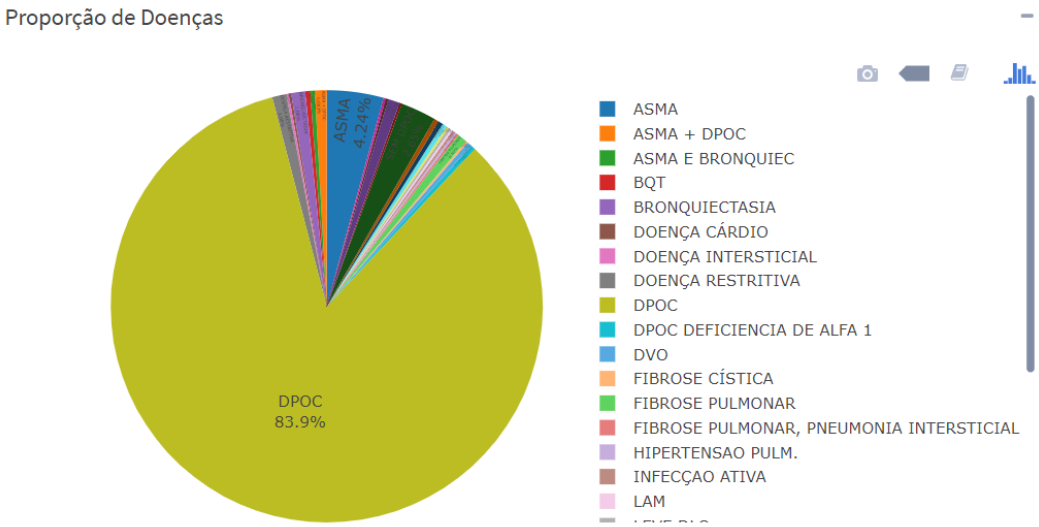
Figura 27 - Percentual geral de abandono.



Fonte: elaborado pelo autor

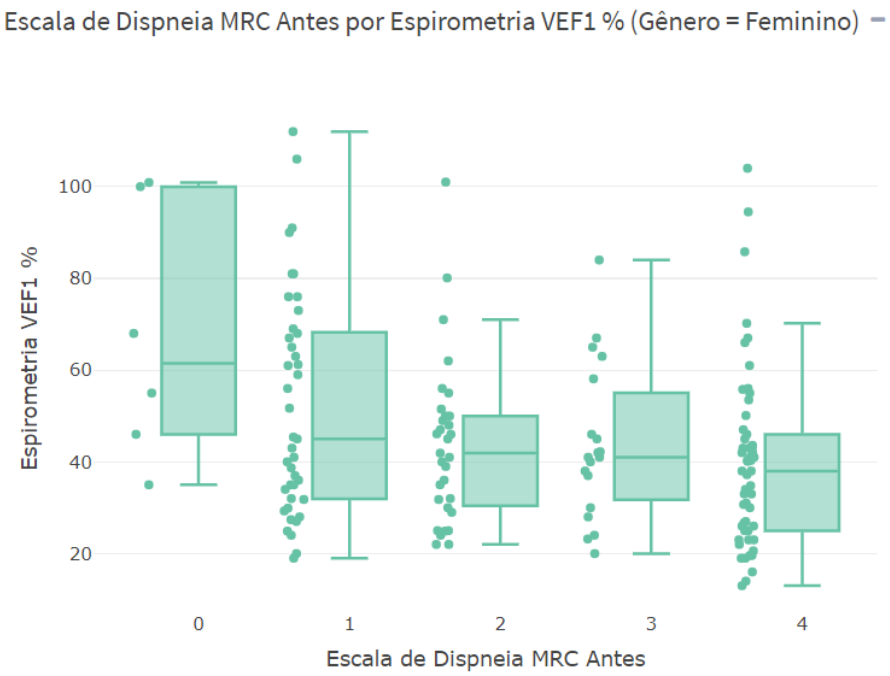
A segunda visualização apresenta a distribuição de doenças da base de dados. Pode-se observar que existe um grande percentual de pacientes portadores de DPOC no projeto de extensão, sendo essa a principal doença que atinge os pacientes do projeto. O percentual de 83,9% equivale a 475 pacientes. Asma aparece em segundo lugar, com 4,24%, totalizando 24 pacientes. A Figura 28 ilustra essa distribuição.

Figura 28 - Proporção de doenças do Projeto de Extensão.



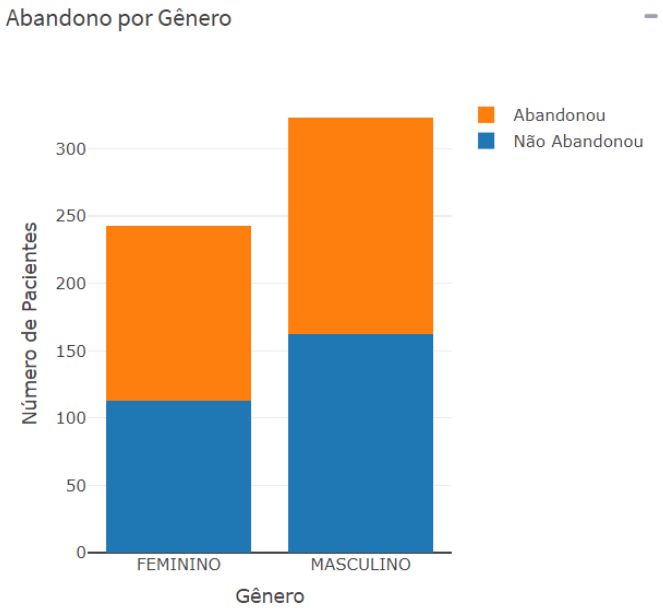
Fonte: elaborado pelo autor

Figura 29 - Escala de dispneia por espirometria VEF1 % no gênero feminino.



Fonte: elaborado pelo autor

Figura 30 - Quantidade de abandonos por gênero.



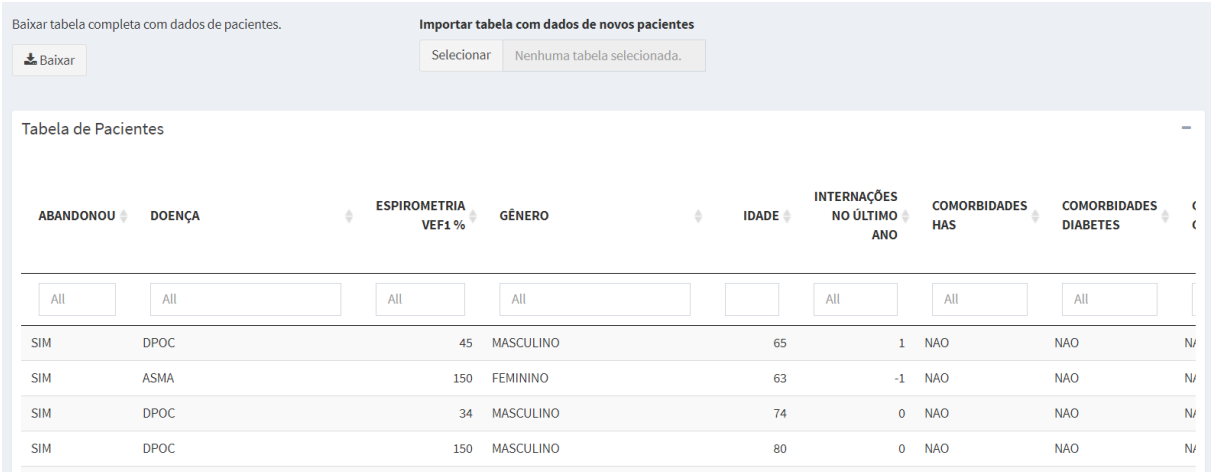
Fonte: elaborado pelo autor

A Figura 29 exibe a relação entre a escala de dispneia e o percentual de espirometria VEF1. Neste gráfico, podemos observar que, conforme a escala de dispneia aumenta, existe uma tendência de redução no percentual de espirometria VEF1, embora existam algumas exceções. A Figura 30 demonstra a quantidade de abandonos por gênero. Pode-se constatar

que, em geral, existe um equilíbrio entre o abandono de pacientes do gênero feminino e masculino. Também é possível verificar que o projeto de extensão recebe um número maior de pacientes do gênero masculino.

A guia “Base de Dados” mostra a tabela completa de pacientes da base de dados do projeto de extensão. Através dessa tabela, é possível realizar filtros e ordenações para localizar determinados registros da base. Nessa guia, também estão disponíveis as funcionalidades de baixar a tabela completa de pacientes em um arquivo com extensão “csv” e importar uma tabela com pacientes que concluíram ou abandonaram o tratamento e ainda não estão na base de dados. A tabela importada deve estar em um arquivo com extensão “csv” e deve possuir os mesmos atributos utilizados para a modelagem apresentada na Subseção 5.1.1. Após a importação da tabela, os novos registros passam a ser considerados para a geração dos modelos preditivos utilizados pela ferramenta, bem como para a geração dos gráficos estatísticos exibidos na guia “Estatísticas”. A Figura 31 demonstra a estrutura da guia “Base de Dados” da ferramenta.

Figura 31 - Guia que apresenta a base de dados do Projeto de Extensão.



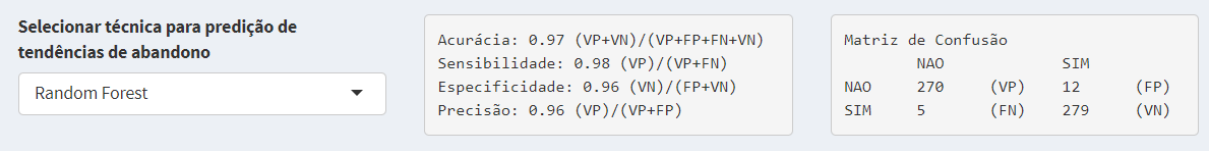
Fonte: elaborado pelo autor

6.1 INVESTIGAÇÃO DO EXCELENTE RESULTADO DE RANDOM FOREST

As métricas de validação de desempenho exibidas na ferramenta são geradas com base no teste dos modelos preditivos sobre o próprio conjunto de treinamento. Em geral, o desempenho dos modelos tende a aumentar nessa situação. No entanto, o modelo preditivo gerado com a utilização da técnica RF demonstrou resultado bastante superior aos resultados apresentados na validação dos modelos gerados com a utilização dessa técnica na Seção 5.3.4.

As métricas exibidas atingiram quase 100% de acerto, diferentemente dos demais modelos. Esse fato motivou a realização de um experimento com o objetivo de validar os resultados apresentados por esse modelo. O desempenho dele é apresentado na Figura 32.

Figura 32 - Métricas de validação de desempenho do modelo RF utilizado pela ferramenta.



Fonte: elaborado pelo autor

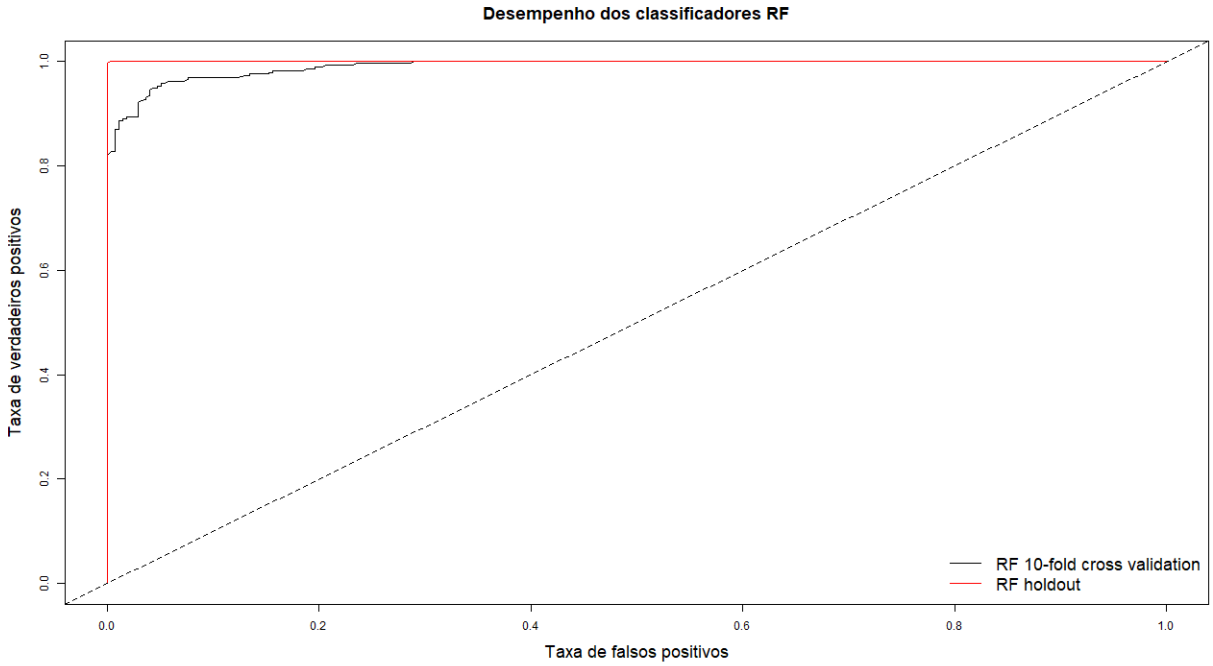
O experimento atestou o comportamento diferente nos modelos gerados com a técnica RF. Ao treinar e validar esses modelos com o mesmo conjunto de dados, o desempenho deles aumenta significativamente, atingindo quase 100% de acerto. Foi analisado o desempenho de dois modelos preditivos gerados com a técnica RF, um treinado com a técnica *10-fold cross validation* e outro treinado uma vez. Para treinamento e teste dos modelos, foi utilizado o conjunto de dados completo, que contém todos os pacientes atendidos pelo projeto de extensão e todos os atributos especificados na Seção 5.1.1. Ambos os modelos obtiveram excelente desempenho. O modelo treinado uma vez classificou corretamente todos os registros da classe positiva, alcançando o valor 1 na sensibilidade. Nas demais métricas, esse modelo aproximou-se do valor 1. O modelo treinado com validação cruzada obteve desempenho um pouco abaixo do modelo treinado uma vez, mas ainda assim o modelo se mostrou bastante eficiente. O Quadro 14 apresenta os resultados obtidos pelos modelos. A Figura 33 exibe as curvas ROC dos modelos analisados.

Quadro 14 - Comparação de desempenho dos modelos preditivos RF treinados e validados com o mesmo conjunto de dados.

Modelo	Acurácia	Sensibilidade	Especificidade	Precisão	AUC
RF 10-fold cross validation	0.9452	0.9709	0.9210	0.9207	0.9905
RF holdout	0.9982	1	0.9966	0.9964	0.9999

Fonte: elaborado pelo autor

Figura 33 - Curvas ROC dos modelos preditivos RF treinados e validados com o mesmo conjunto de dados.



Fonte: elaborado pelo autor

6.2 VALIDAÇÃO DA FERRAMENTA

Após a conclusão do desenvolvimento da ferramenta, foi realizada a validação da mesma com a professora coordenadora do projeto de extensão. Foram elaboradas 5 questões com o objetivo de verificar a qualidade e a relevância do conhecimento gerado sobre o projeto. As perguntas e respostas seguem abaixo:

1. A qualidade de interface dos gráficos está de acordo?

R: Sim, os gráficos ficaram muito bons.

2. A qualidade do conhecimento gerado está de acordo?

R: Sim, visualizamos informações muito importantes.

3. Os modelos preditivos ficaram intuitivos?

R: Ficaram sim.

4. Os modelos preditivos trouxeram informações relevantes sobre o abandono?

R: Com certeza. Além disso, foram incluídas muitas informações que poderemos explorar com o passar do tempo.

5. Sugestões de trabalhos futuros.

R: É necessário um tempo maior de familiarização com o instrumento para, assim, elaborarmos juntos outros instrumentos.

Com a validação é possível observar que a ferramenta apresentou tanto qualidade visual como de conhecimentos gerados. Os modelos preditivos intuitivos facilitam a compreensão dos resultados, não exigindo grande conhecimento técnico em informática para isso. Por fim, além de informações relevantes sobre o abandono, as visualizações geradas pela ferramenta criaram a oportunidade de exploração de informações que não são estudadas atualmente pela dificuldade de análise dos dados.

6.3 CONSIDERAÇÕES FINAIS

Finalizada a comparação entre os modelos, foi iniciado o desenvolvimento da ferramenta para aplicação das técnicas de análise preditiva e geração de visualizações sobre os dados do projeto de extensão. A ferramenta foi desenvolvida com a linguagem de programação R. O painel de visualização foi dividido em três guias. A primeira possibilita a análise de pacientes que estão ingressando no tratamento, bem como, apresenta gráficos sobre a distribuição dos pacientes analisados. A segunda guia apresenta gráficos estatísticos sobre todos os registros e atributos do conjunto de dados, visando ilustrar uma estatística descritiva sobre os dados do projeto. Nessas visualizações, são considerados apenas os pacientes que concluíram ou abandonaram o tratamento. A terceira guia apresenta o conjunto de dados completo, possibilitando a aplicação de filtros e ordenações. Nessa guia, também é possível baixar a tabela completa de pacientes e importar registros de pacientes que finalizaram ou abandonaram o tratamento e ainda não estão armazenados no conjunto de dados.

Essa ferramenta foi avaliada pela professora coordenadora do projeto de extensão após a conclusão do desenvolvimento. A validação foi realizada por meio de um questionário com 5 perguntas e demonstrou que o conhecimento gerado pela ferramenta é de grande importância para o projeto. A qualidade visual dos gráficos interativos também foi atestada. Por fim, outro ponto positivo da ferramenta foi a criação da oportunidade de exploração de informações que não são analisadas atualmente.

7 CONCLUSÃO

Essa pesquisa apresenta o “Projeto de Extensão Reabilitação Pulmonar” da Universidade Feevale cujos dados foram utilizados para aplicação de análise preditiva. A pesquisa identificou oportunidades de utilização de *machine learning* nesse projeto de extensão, bem como alguns benefícios que o uso dessa tecnologia pode gerar.

Inicialmente, foi realizada uma revisão sistemática para identificação de técnicas de análise preditiva aplicadas à reabilitação pulmonar. Na revisão, também foi realizado o levantamento de linguagens de programação, técnicas, ferramentas e formas de validação mais utilizadas nessa área. Com base na revisão sistemática realizada, a linguagem de programação R foi selecionada para auxiliar a análise preditiva. Além disso, as técnicas de análise preditiva (Seção 5.2) e as formas de validação (Seção 5.3) utilizadas neste trabalho também foram selecionadas com base na revisão.

As técnicas de análise preditiva SVM, DT e RF foram aplicadas sobre os dados do projeto de extensão, visando prever a tendência de abandono dos pacientes que estão ingressando no tratamento. Um dos modelos preditivos gerados com a aplicação da técnica RF atingiu melhor desempenho entre os modelos analisados. Com essa análise, observou-se que a aplicação de análise preditiva pode auxiliar na previsão de tendência de abandono dos pacientes. A partir dessa análise, é possível traçar estratégias para reduzir o grande percentual de abandono do tratamento.

A ferramenta desenvolvida possibilita que especialistas da área analisem a tendência de abandono de um grupo de pacientes sem necessidade de muito conhecimento na área de informática. A ferramenta disponibiliza um painel interativo que apresenta de forma amigável o resultado da análise do grupo de pacientes que está iniciando o tratamento. Também disponibiliza visualizações sobre o grupo de pacientes analisado, apresentando a distribuição deles. Além disso, visualizações são geradas sobre todos os registros e atributos do conjunto de dados com o objetivo de gerar estatísticas relevantes para o projeto.

A fim de verificar a qualidade da ferramenta desenvolvida, foi realizada a validação da mesma pela professora coordenadora do projeto de extensão por meio de um questionário com 5 perguntas. Através da validação, é possível concluir que a ferramenta apresenta qualidade visual nos gráficos gerados. Além disso, também foi afirmada a qualidade das informações geradas pela ferramenta, que fornecem conhecimento relevante para o projeto.

Também foi sinalizado como ponto positivo a exibição de informações relevantes que atualmente não são avaliadas, mas podem ser exploradas futuramente com o auxílio da ferramenta.

Durante o desenvolvimento deste trabalho, foram identificadas algumas limitações. O conjunto de dados utilizados para análise contém somente 566 registros. Essa quantidade é pequena para a análise preditiva. Uma quantidade maior de registros faria uma generalização mais coerente e melhoraria a identificação de padrões internos do conjunto de dados. Consequentemente, aumentaria o desempenho dos modelos preditivos. Além disso, a grande quantidade de pacientes portadores de DPOC ocasionou o desbalanceamento do conjunto de dados. Essa situação também impacta no desempenho dos modelos apesar de ter sido utilizada uma estratégia para geração de um subconjunto balanceado para treinamento dos modelos.

O armazenamento dos dados em uma planilha Excel também ocasionou alguns problemas na coleta dos dados. Um dos problemas identificados foi a falta de padronização de doenças, tornando necessário esse ajuste antes de iniciar a análise dos dados. Outros problemas gerados por essa forma de coleta são os problemas de digitação e a falta de registro de determinado atributo pela não obrigatoriedade de preenchimento.

O abandono do tratamento por pacientes também resulta na falta de coleta de determinados atributos. A ausência de valores impacta no desempenho dos modelos. Além disso, as técnicas SVM e RF exigem um conjunto de dados sem valores ausentes. Neste trabalho, foram adotadas estratégias para preenchimento de valores ausentes. No entanto, valores reais gerariam melhor desempenho dos modelos preditivos, visto que facilitariam a identificação de padrões do conjunto de dados. Uma grande limitação nesse contexto de valores ausentes foi o atributo “Escala de Dispnéia MRC_Antes”, que possuía 215 valores ausentes, equivalentes a boa parte dos registros do conjunto de dados. Essa grande ausência de valores motivou a criação de modelos preditivos desconsiderando esse atributo, mas que se mostraram menos eficientes.

Em relação aos modelos preditivos utilizados pela ferramenta, foi identificada uma limitação para visualização gráfica do modelo criado com a técnica SVM. A utilização de múltiplos atributos inviabilizou a visualização do modelo graficamente, visto que não é possível gerar visualizações com mais de três dimensões. Além disso, diferentemente do que ocorreu com o modelo criado com a técnica RF, não foi identificada nenhuma técnica para visualização dos atributos mais importantes para o modelo.

Como continuidade deste trabalho, evoluções na análise podem ser realizadas. Em relação aos modelos preditivos, mais registros podem ser utilizados para a análise, visando validar os resultados obtidos. Além disso, pode ser explorada a inclusão de novos atributos e novas técnicas de análise preditiva. Em relação à ferramenta desenvolvida, podem ser disponibilizados novos gráficos e evoluída a importação da tabela de novos pacientes que, atualmente, é realizada por meio de uma planilha Excel. Outro trabalho futuro importante para o contexto do projeto e a qualidade dos dados é o desenvolvimento de um sistema para entrada e gerenciamento dos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

ALPAYDIN, Ethem. **Introduction to Machine Learning**. 2° ed. Cambridge: The MIT Press, 2010.

AMARAL, Jorge L. M.; LOPES, Agnaldo J.; FARIA, Alvaro C. D.; MELO, Pedro L. Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease. **Computer Methods and Programs in Biomedicine**, v. 118, n. 2, p. 186–197, 2015.

AMARAL, Jorge L. M.; LOPES, Agnaldo J.. VEIGA, Juliana; FARIA, Alvaro C. D.; MELO, Pedro L. High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements. **Computer Methods and Programs in Biomedicine**, v. 144, p. 113–125, 2017.

AMERICAN ASSOCIATION OF CARDIOVASCULAR AND PULMONARY REHABILITATION. **Diretrizes para programas de reabilitação pulmonar: promovendo a saúde e prevenindo a doença**. 3° ed. São Paulo: Roca, 2007.

BARRETO, Sérgio S. Menna. Volumes Pulmonares. **Jornal Brasileiro de Pneumologia**, v. 28, n. 3, p. S83–S94, 2002.

BAUERLE, Otto; CHRUSCH, Carla A.; YOUNES, Magdy. Mechanisms by which COPD affects exercise tolerance. **American Journal of Respiratory and Critical Care Medicine**, v. 157, n. 1, p. 57–68, 1998.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CHE, Zhengping.; PURUSHOTHAM, Sanjay; KHEMANI, Robinder; LIU, Yan. **Interpretable Deep Models for ICU Outcome Prediction.** AMIA ... Annual Symposium proceedings. AMIA Symposium. **Anais** ... 2016. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/28269832%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5333206>

COSTA, C. C. da; LERMEN, C. de Azeredo; COLOMBO, C.; CANTERLE, D. B.; MACHADO, M. L. L.; KESSLER, A.; TEIXEIRA, P. J. Z. Effect of a Pulmonary Rehabilitation Program on the levels of anxiety and depression and on the quality of life of patients with chronic obstructive pulmonary disease. **Revista Portuguesa de Pneumologia**, v. 20, n. 6, p. 299–304, 2014.

COSTA, D.; JAMAMI, M. Bases fundamentais da espirometria. **Rev. bras. fisioter.**, v. 5, n. 2, p. 95–102, 2001.

DAS, Nilakash; TOPALOVIC, Marko; JANSSENS, Wim. Artificial intelligence in diagnosis of obstructive lung disease: Current status and future potential. **Current Opinion in Pulmonary Medicine**, v. 23, n. 00, p. 1–7, 2017.

DeCS - Descritores em Ciências da Saúde. Disponível em: <http://decs.bvs.br>. Acesso em: 17 maio 2018.

DUCHIADE, Milena P. Poluição do ar e doenças respiratórias: uma revisão. **Cadernos de Saúde Pública**, v. 8, n. 3, p. 311–330, 1992.

EDHLUND, Bengt M.; MCDOUGALL, Allan G. **PubMed Essentials, Mastering the World's Health Research Database**. 3º ed. Lulu.com, 2014.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André C. P. L. F. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro: LTC, 2011.

FERNANDEZ-GRANERO, Miguel Angel; SANCHEZ-MORILLO, Daniel; LOPEZ-GORDO, Miguel Angel; LEON, Antonio. **A Machine Learning Approach to Prediction of Exacerbations of Chronic Obstructive Pulmonary Disease**International Work-Conference on the Interplay Between Natural and Artificial Computation. **Anais**...Elche: Springer, 2015. Disponível em: <http://link.springer.com/10.1007/978-3-319-18914-7>

FERREIRA, Juliana Carvalho; PATINO, Cecilia Maria. Entendendo os testes diagnósticos. Parte 1. v. 43, n. 5, p. 330, 2017.

FILHO, João. Avaliação laboratorial da função Pulmonar. **Medicina (Ribeirao Preto. Online)**, v. 31, n. 2, p. 191–207. In: SIMPÓSIO DE DOENÇAS PULMONARES. Ri, 1998.

FINKELSTEIN, Joseph; JEONG, In cheol. Machine learning approaches to personalize early prediction of asthma exacerbations. **Ann N Y Acad Sci**, v. 1387, n. 1, p. 153–165, 2017.

FRANCESCHET, Massimo. A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. **Scientometrics**, v. 83, n. 1, p. 243–258, 2010.

GILMAN, S. A.; BANZETT, R. B. Physiologic changes and clinical correlates of advanced dyspnea. **Current Opin. in Support palliative Care**, v. 3, n. 2, p. 93–97, 2009.

GIRALDO, Sebastián Robledo; ZULUAGA, German Augusto Osorio; ESPINOSA, Carolina López. Networking en pequeña empresa: una revisión bibliográfica utilizando la teoria de grafos. **Revista Vinculos**, v. 11, n. 2, p. 6–16, 2014.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining: Um Guia Prático**. Rio de Janeiro: Elsevier Editora Ltda., 2005.

GOULART, Flavio A. de Andrade. Doenças crônicas não transmissíveis: estratégias de controle e desafios e para os sistemas de saúde. **Organização Pan-Americana da Saúde/ Organização Mundial da Saúde**, p. 96, 2011.

GUAN, W-J; JIANG, M; GAO, Y-H; LI, H-M; XU, G; ZHENG, J-P; CHEN, R-C; ZHONG, N-S. Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics. **Int J Tuberc Lung Dis**, v. 20, n. 3, p. 402–410, 2016.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques**. 2° ed. San Francisco: Elsevier Inc., 2006.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3° ed. Waltham: Elsevier, 2012.

VAN DER HEIJDEN, Maarten; VELIKOVA, Marina; LUCAS, Peter J.F.. Learning Bayesian networks for clinical time series analysis. **Journal of Biomedical Informatics**, v. 48, p. 94–105, 2014.

HOLCZER, Balazs. **Random Forest Classifier – Machine Learning**. Disponível em: <http://www.globalsoftwaresupport.com/random-forest-classifier-bagging-machine-learning/>. Acesso em: 5 jun. 2018.

JENNINGS, Jeffrey H.; DIGIOVINE, Bruno; OBEID, Dany; FRANK, Cathy. The association between depressive symptoms and acute exacerbations of COPD. **Lung**, v. 187, n. 2, p. 128–135, 2009.

KITCHENHAM, Elisabete. **Guidelines for performing Systematic Literature Reviews in Software Engineering version 2**. Durham: Department of Computer Science University of Durham Durham, UK, 2007.

LAPES. **StArt.** São Carlos. Computing Department of the Federal University of São Carlos (DC/UFSCar), 2018. Disponível em: <http://lapes.dc.ufscar.br/tools/start_tool>. Acesso em: 26 maio. 2018.

LEIDY, Nancy K.; MALLEY, Karen G.; STEENROD, Anna W.; MANNINO, David M.; MAKE, Barry J.; BOWLER, Russ P.; THOMASHOW, Byron M.; BARR, R. G.; RENNARD, Stephen I.; HOUFEK, Julia F.; YAWN, Barbara P.; HAN, Meilan K.; MELDRUM, Catherine A.; BACCI, Elizabeth D.; WALSH, John W.; MARTINEZ, Fernando. Insight into Best Variables for COPD Case Identification: A Random Forests Analysis. **Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation**, v. 3, n. 1, p. 406–418, 2016.

LIU, Bing. **Web Data Mining: Exploring Hyperlinks, Contents and Usage Data**. 2° ed.

Chicago: Springer, 2011.

LUO, Gang; NKOY, Flory L.; STONE, Bryan L.; SCGMICK, Darell; JOHNSON, Michael D. A systematic review of predictive models for asthma development in children. **BMC Medical Informatics and Decision Making**, v. 15, n. 99, p. 1–16, 2015.

MALTA, Deborah Carvalho; MOURA, Lenildo de; PRADO, Rogério Ruscitto do; ESCALANTE, Juan Cortez; SCHMIDT, Maria Inês; DUNCAN, Bruce Bartholow. Mortalidade por doenças crônicas não transmissíveis no Brasil e suas regiões, 2000 a 2011. **Epidemiologia e Serviços de Saúde**, v. 23, n. 4, p. 599–608, 2014.

MAURER, Janet; REBBAPRAGADA, Venkata; BORSON, Soo; GOLDSTEIN, Roger; KUNIK, Mark E.; YOHANNES, Abebaw M.; HANANIA, Nicola A. Anxiety and depression in COPD: Current understanding, unanswered questions, and research needs. **Chest**, v. 134, n. 4, p. 43S–56S, 2008.

MEDEIROS, Edna Ramos de. **Revisão Sistemática Sobre os Dispositivos Vestíveis na Área da Saúde**. Universidade Feevale, 2016.

MITCHELL, Tom M. The Discipline of Machine Learning. 2006.

MITRA, Sushmita; ACHARYA, Tinku. **Data Mining: Multimedia, Soft Computing, and Bioinformatics**. New Jersey: Wiley, 2003.

NICI, Linda; DONNER, Claudio; WOUTERS, Emiel; ZUWALLACK, Richard; AMBROSINO, Nicolino; BOURBEAU, Jean; CARONE, Mauro; CELLI, Bartolome; ENGELEN, Marielle; FAHY, Bonnie; GARVEY, Chris; GOLDSTEIN, Roger; GOSSELINK, Rik; LAREAU, Suzanne; MACINTYRE, Neil; MALTAIS, Francois; MORGAN, Mike; O'DONNELL, Denis; PREFAULT, Christian; REARDON, Jane; ROCHESTER, Carolyn; SCHOLS, Annemie; SINGH, Sally; TROOSTERS, Thierry.. American Thoracic Society/European Respiratory Society Statement on Pulmonary Rehabilitation. **American Journal of Respiratory and Critical Care Medicine**, v. 173, n. 1, p. 1390–1413, 2006.

PALOMBINI, Bruno Carlos; PORTO, Nelson da Silva; ARAÚJO, Elisabeth; GODOY, Dagoberto Vanoni de. **Doenças das vias aéreas: uma visão clínica integradora (viaerologia)**. Rio de Janeiro: Revinter, 2001.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent; VANDERPLAS, Jake; PASSOS, Alexandre; COURNAPEAU, David; BRUCHER, Matthieu; PERROT, Matthieu; DUCHESNAY,

Édouard. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PELLEGRINO, R.; VIEGI, G.; BRUSASCO, Vito; CRAPO, R. O.; BURGOS, F.; CASABURI, R.; COATES, A.; VAN DER GRINTEN, C. P M; GUSTAFSSON, P.; HANKINSON, J.; JENSEN, R.; JOHNSON, D. C.; MACINTYRE, N.; MCKAY, R.; MILLER, M. R.; NAVAJAS, D.; PEDERSEN, O. F.; WANGER, J. Interpretative strategies for lung function tests. **European Respiratory Journal**, v. 26, n. 5, p. 948–968, 2005.

PIEDRA, David; FERRER, Antoni; GEA, Joaquim. Text Mining and Medicine: Usefulness in Respiratory Diseases. **Archivos de Bronconeumología (English Edition)**, v. 50, n. 3, p. 113–119, 2014.

POHAR, Maja; BLAS, Mateja; TURK, Sandra. Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. **Metodološki zvezki**, v. 1, n. 1, p. 143–161, 2004.

POULAIN, Magali; DURAND, Fabienne; PALOMBA, Bernard; CEUGNIET, François; DESPLAN, Jacques; VARRAY, Alain; PRÉFAUT, Christian. 6-Minute walk testing is more sensitive than maximal incremental cycle testing for detecting oxygen desaturation in patients with COPD. **Chest**, v. 123, n. 5, p. 1401–1407, 2003.

PRACTICALAI ONLINE. **Implementing Classification using Logistic Regression in Ruby**. Disponível em: <https://www.practicalai.io/implementing-classification-using-logistic-regression-in-ruby/>. Acesso em: 7 jun. 2018.

PRISMA. **Preferred reporting items for Systematic Reviews and Meta-Analyses**. Disponível em: <http://prisma-statement.org/>. Acesso em: 26 maio. 2018.

Pubmed. Disponível em: <https://www.ncbi.nlm.nih.gov/pubmed/>. Acesso em: 26 maio. 2018.

PYTHON SOFTWARE FOUNDATION. **Python**. Disponível em: <https://www.python.org>. Acesso em: 9 jun. 2018.

REGUEIRO, Eloisa Maria Gatti; LORENZO, Valéria Amorim Pires Di; PARIZOTTO, Ana Paula De Deus; NEGRINI, Fernanda; SAMPAIO, Luciana Maria Malosá. Análise da demanda metabólica e ventilatória durante a execução de atividades de vida diária em indivíduos com doença pulmonar obstrutiva crônica. **Revista latino-americana de enfermagem**, v. 14, n. 1, p. 41–7, 2006.

RIES, Andrew L.; BAULDOFF, Gerene S.; CARLIN, Brian W.; CASABURI, Richard; EMERY, Charles F.; MAHLER, Donald A.; MAKE, Barry; ROCHESTER, Carolyn L.;

ZUWALLACK, Richard; HERRERIAS, Carla. Pulmonary rehabilitation: Joint ACCP/AACVPR evidence-based guidelines. **Chest**, v. 131, n. 5, p. 4S–42S, 2007.

RODRIGUES, Joaquim Carlos; CARDIERI, Joselina M. Andrade; BUSSAMRA, Maria Helena Carvalho de Ferreira; NAKAIE, Cleyde Myriam Aversa; ALMEIDA, Marina Buarque de; FILHO, Luiz Vicente Ferreira da Silva; ADDE, Fabíola Villac. Provas de função pulmonar em crianças e adolescentes. **Jornal Pneumologia**, v. 28, n. Supl 3, p. S 207-S 221, 2002.

RSTUDIO. **RStudio**. Disponível em: <https://www.rstudio.com/products/rstudio/>. Acesso em: 9 jun. 2018.

SANCHEZ-MORILLO, Daniel; FERNANDEZ-GRANERO, Miguel A.; LEON-JIMENEZ, Antonio. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review. **Chronic Respiratory Disease**, v. 13, n. 3, p. 264–283, 2016.

SECRETARIA DE VIGILÂNCIA EM SAÚDE − MINISTÉRIO DA SAÚDE. Epidemiológico. **Boletim Epidemiológico**, v. 47, n. 19, p. 1–9, 2016.

SHAH, Syed Ahmar; VELARDO, Carmelo; FARMER, Andrew; TARASSENKO, Lionel. Exacerbations in chronic obstructive pulmonary disease: Identification and prediction using a digital health system. **Journal of Medical Internet Research**, v. 19, n. 3, p. 1–14, 2017.

SILVA, Fernando Henrique da. **Estudo e desenvolvimento de métodos para predição de doadores de sangue.** 2018. 80f. Dissertação (Mestrado em Modelagem e Otimização) - Universidade Federal de Goiás, Catalão, GO, 2018.

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à Mineração de Dados com aplicações em R**. Rio de Janeiro: Elsevier, 2016.

SOCIEDADE PAULISTA DE PNEUMOLOGIA E TISIOLOGIA. **Pneumologia: atualização e reciglagem**. 7.ed. ed. São Paulo: Roca, 2008.

SOCIEDADE BRASILEIRA DE PNEUMOLOGIA E TISIOLOGIA. II Consenso Brasileiro sobre Doença Pulmonar Obstrutiva Crônica - DPOC. **Jornal Brasileiro de Pneumologia**, v. 30, n. 5, p. S1–S42, 2004.

SPATHIS, Dimitris; VLAMOS, Panayiotis. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. **Health Informatics Journal**, p. 1–17, 2017.

SPERANDIO, Evandro Fornias; ARANTES, Rodolfo Leite; MATHEUS, Agatha Caveda; PEREIRA, Rodrigo; LAURIA, Vinícius Tonon; ROMITI, Marcello; RICARDO, Antônio; GAGLIARDI, De Toledo; DOURADO, Victor Zuniga.. Distúrbio ventilatório restritivo

sugerido por espirometria : associação com risco cardiovascular e nível de atividade física em adultos assintomáticos. **Jornal Brasileiro Pneumologia**, v. 42, n. 1, p. 22–28, 2016.

SU, Chong; JU, Shenggen; LIU, Yiguang; YU, Zhonghua. An Empirical Study of Skew-Insensitive Splitting Criteria and its Application in Traditional Chinese Medicine. **Intelligent Automation and Soft Computing**, v. 20, n. 4, p. 535–554, 2014.

SWAMINATHAN, Sumanth; QIRKO, Klajdi; SMITH, Ted; CORCORAN, Ethan; WYSHAM, Nicholas G.; BAZAZ, Gaurav; KAPPEL, George; GERBER, Anthony N. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. **PLoS ONE**, v. 12, n. 11, p. 1–21, 2017.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining**. Rio de Janeiro: Editora Ciência Moderna, 2009.

TARANTINO, Affonso Berardinelli. **Doenças Pulmonares**. 5° ed. Rio de Janeiro: GUANABARA KOOGAN, 2002.

THE R FOUNDATION. **The R Project for Statistical Computing**. Disponível em: <https://www.r-project.org/>. Acesso em: 9 jun. 2018.

TOPALOVIC, Marko; EXADAKTYLOS, Vasileios; DECRAMER, Marc; TROOSTERS, Thierry; BERCKMANS, Daniel; JANSSENS, Wim. Modelling the dynamics of expiratory airflow to describe chronic obstructive pulmonary disease. **Medical & Biological Engineering & Computing**, v. 52, n. 12, p. 997–1006, 2014.

TOPALOVIC, Marko; LAVAL, Stefan; AERTS, Jean Marie; TROOSTERS, Thierry; DECRAMER, Marc; JANSSENS, Wim. Automated Interpretation of Pulmonary Function Tests in Adults with Respiratory Complaints. **Respiration**, v. 93, n. 3, p. 170–178, 2017.

VIDALE, G. **44% dos brasileiros sofrem com problemas respiratórios**. Disponível em: <https://veja.abril.com.br/saude/44-dos-brasileiros-sofrem-com-problemas-respiratorios/>. Acesso em: 17 mar. 2018.

VIEIRA, Paula Vanessa Medeiros; WAINER, Jacques. Correlações entre a contagem de citações de pesquisadores brasileiros, usando o Web of Science, Scopus e Scholar. **Perspectivas em Ciência da Informação**, v. 18, n. 3, p. 45–60, 2013.

WANG, Qi; WALTMAN, Ludo. Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. **Journal of Informetrics**, v. 10, n. 2, p. 347–364, 2016.

Web of Science. Disponível em: <http://apps-webofknowledge.ez310.periodicos.capes.gov.br/>. Acesso em: 26 maio. 2018.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining: Practical Machine Learning Tools and Techniques**. 2° ed. Burlington: Elsevier Inc., 2005.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. **Data mining: Practical Machine Learning Tools and Techniques**. 3° ed. Burlington: Elsevier Inc., 2011.

WU, Wei; BLEECKER, Eugene; MOORE, Wendy; BUSSE, William W.; CASTRO, Mario; CHUNG, Kian Fan; CALHOUN, William J.; ERZURUM, Serpil; GaSTON, Benjamin; ISRAEL, Elliot; CURRAN-EVERETT, Douglas; WENZEL, Sally E. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. **Journal of Allergy and Clinical Immunology**, v. 133, n. 5, p. 1280–1288, 2014.

ZANCHET, Renata Cláudia; VIEGAS, Carlos Alberto Assis; LIMA, Terezinha. A eficácia da reabilitação pulmonar na capacidade de exercício, força da musculatura inspiratória e qualidade de vida de portadores de doença pulmonar obstrutiva crônica. **Jornal Brasileiro de Pneumologia**, v. 31, n. 2, p. 118–124, 2005.

ZHONG, Xiang; LEE, Sujee; ZHAO, Cong; LEE, Hyo Kyung; BAIN, Philip A.; KUNDINGER, Tammy; SOMMERS, Craig; BAKER, Christine; LI, Jingshan. Reducing COPD readmissions through predictive modeling and incentive-based interventions. **Health Care Management Science**, p. 1–19, 2017.

APÊNDICE A – DICIONÁRIO DOS DADOS ANALISADOS

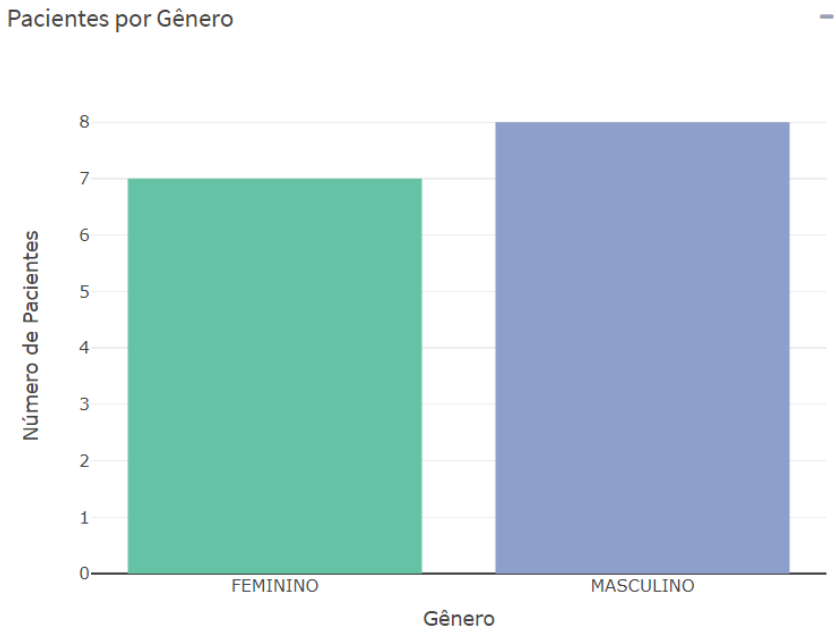
Quadro A-1 - Dicionário dos dados analisados.

Atributo	Descrição
Doença	Doença portada pelo paciente.
Espirometria VEF1%	Percentual do volume expiratório do primeiro segundo medido na espirometria simples. Quanto menor o valor, mais doente o paciente está.
Gênero	Indica o gênero do paciente. 1 - Masculino, 2 - Feminino.
Idade	Idade do paciente.
AC_Quantas internações no último ano?	Número de internações do paciente no último ano.
Comorbidades HAS	Indica se o paciente tem Hipertensão Arterial Sistêmica. 1 - Sim, 2 - Não.
Comorbidades Diabetes	Indica se o paciente tem Diabetes. 1 - Sim, 2 - Não.
Comorbidades Cardiopatias	Indica se o paciente tem Doenças Cardíacas. 1 - Sim, 2 - Não.
Comorbidades Outras	Indica se o paciente tem outras comorbidades. Preenchido - Sim, 2 - Não.
Escala de Dispnéia MRC_Antes	Indica a gravidade da falta de ar. Varia de 0 a 4. Quanto maior o valor, pior o estado do paciente. Essa informação é coletada através do formulário Escala de Dispneia MRC.
Abandonou	Indica se o paciente abandonou o tratamento.
QQVSG Sint.Depois	Indica o resultado do Questionário do Hospital Saint George do paciente após a conclusão do tratamento.
TCSatO2i Depois	Indica a saturação de oxigênio do paciente após a conclusão do tratamento.

Fonte: elaborado pelo autor

APÊNDICE B – VISUALIZAÇÕES DA GUIA “NOVOS PACIENTES” DA FERRAMENTA DESENVOLVIDA

Figura B-1 - Quantidade de pacientes por gênero.

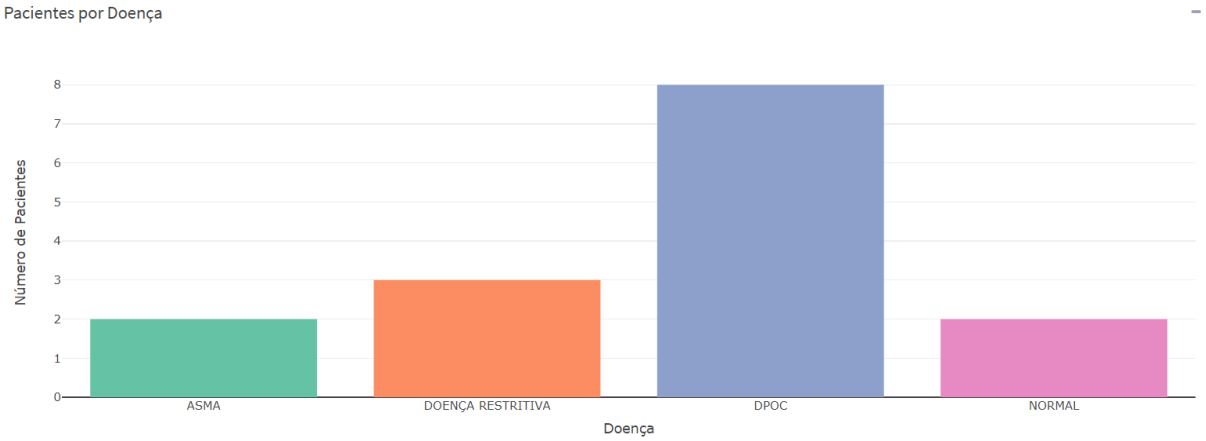


Fonte: elaborado pelo autor

Figura B-2 - Distribuição de idade por gênero.

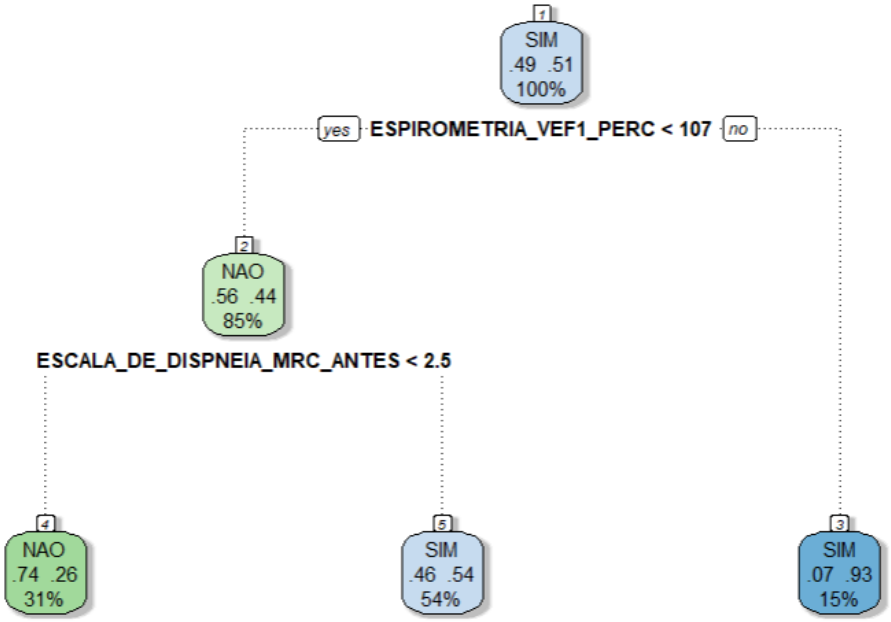
Fonte: elaborado pelo autor

Figura B-3 - Quantidade de pacientes por doença.



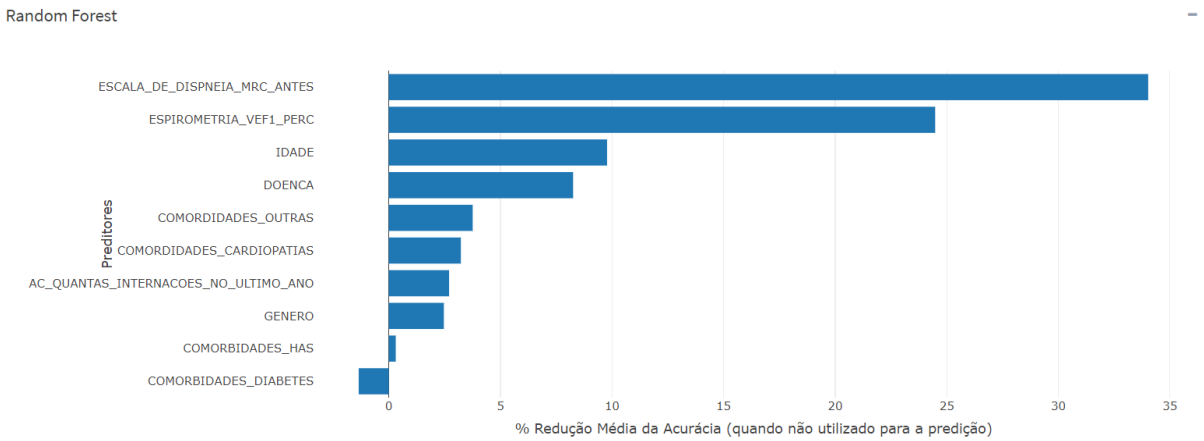
Fonte: elaborado pelo autor

Figura B-4 - Árvore de decisão utilizada atualmente pela ferramenta.



Fonte: elaborado pelo autor

Figura B-5 - Preditores mais importantes para o modelo Random Forest utilizado atualmente pela ferramenta.



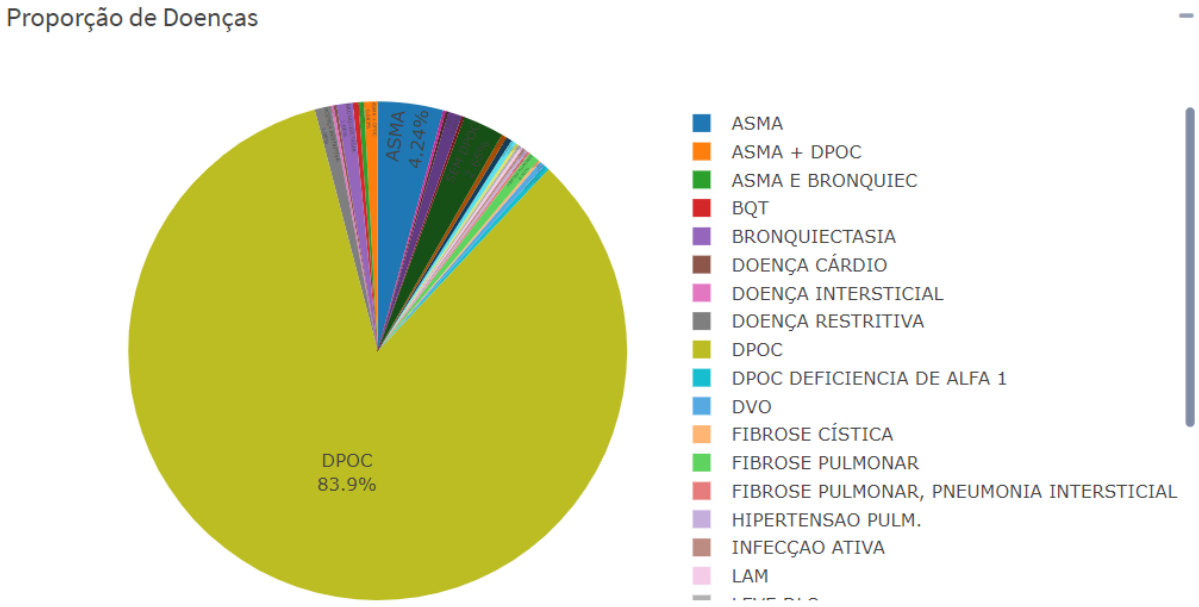
Fonte: elaborado pelo autor

APÊNDICE C – VISUALIZAÇÕES DA GUIA “ESTATÍSTICAS” DA FERRAMENTA DESENVOLVIDA

Figura C-1 - Percentual geral de abandono.

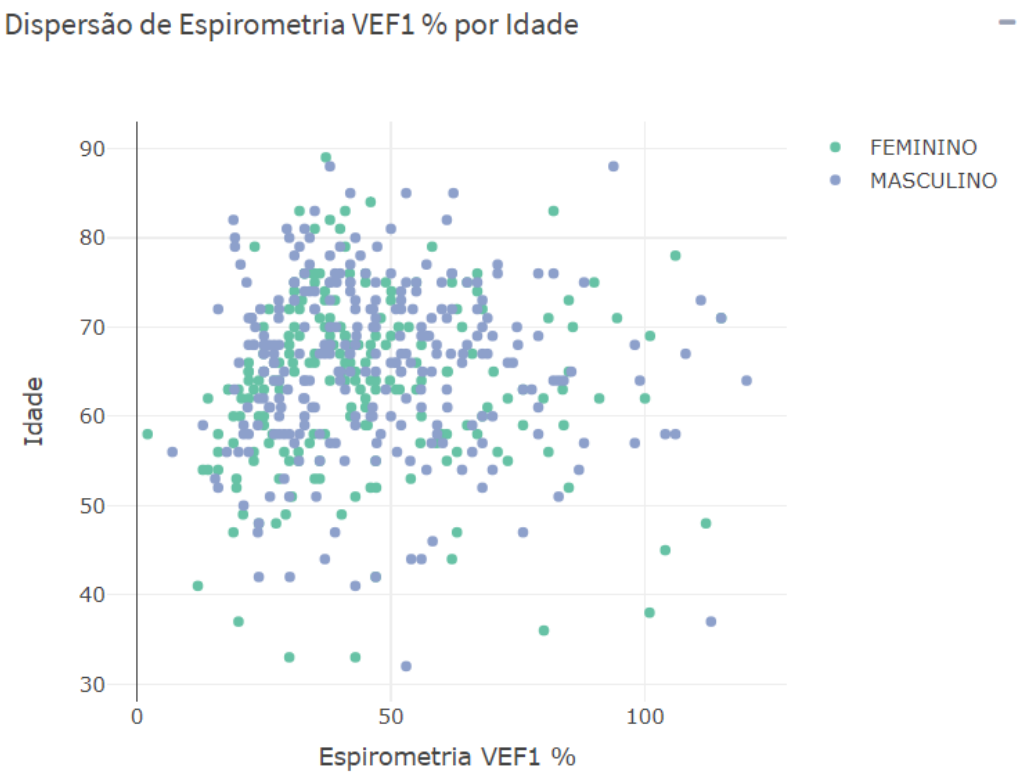
Fonte: elaborado pelo autor

Figura C-2 - Proporção de doenças.



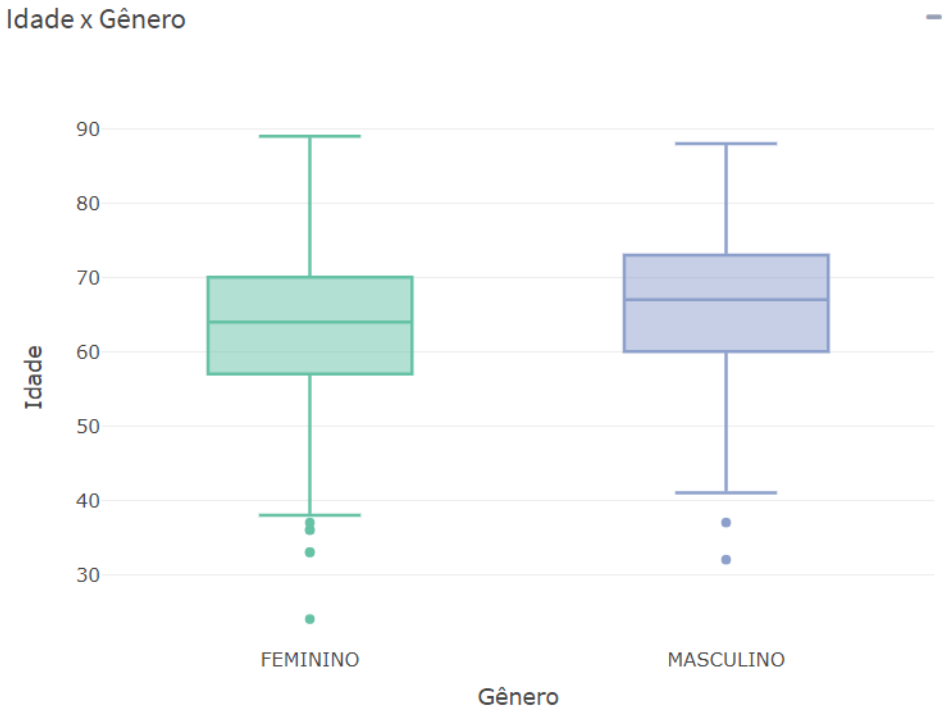
Fonte: elaborado pelo autor

Figura C-3 - Dispersão de espirometria VEF1 % por idade.



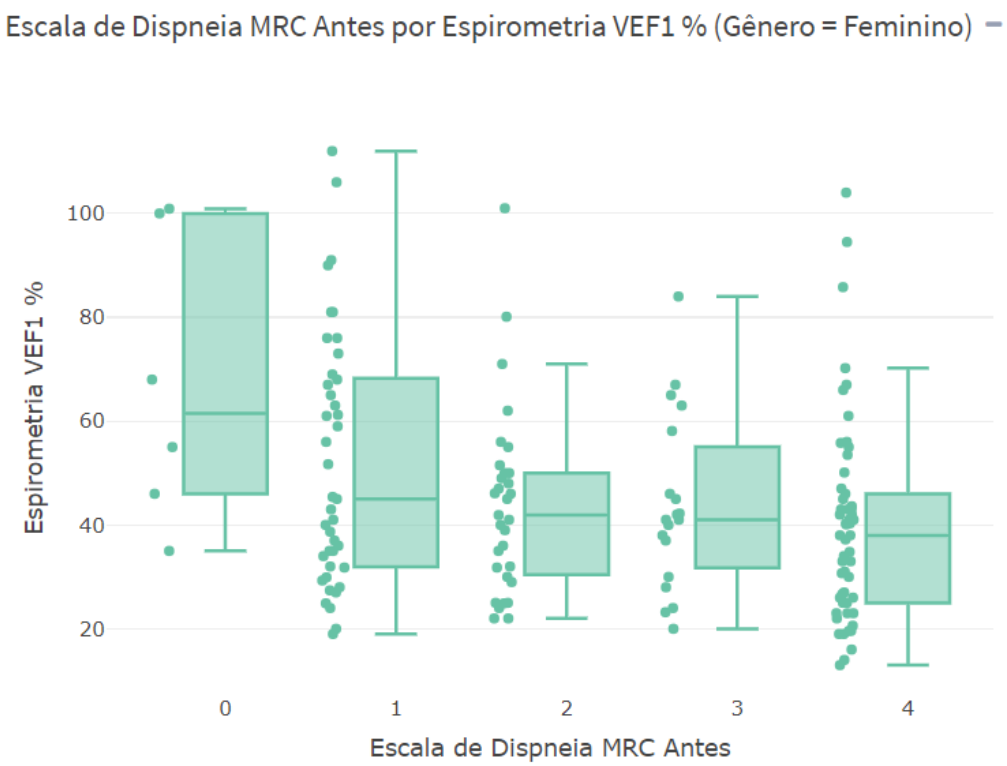
Fonte: elaborado pelo autor

Figura C-4 - Distribuição de idade por gênero.



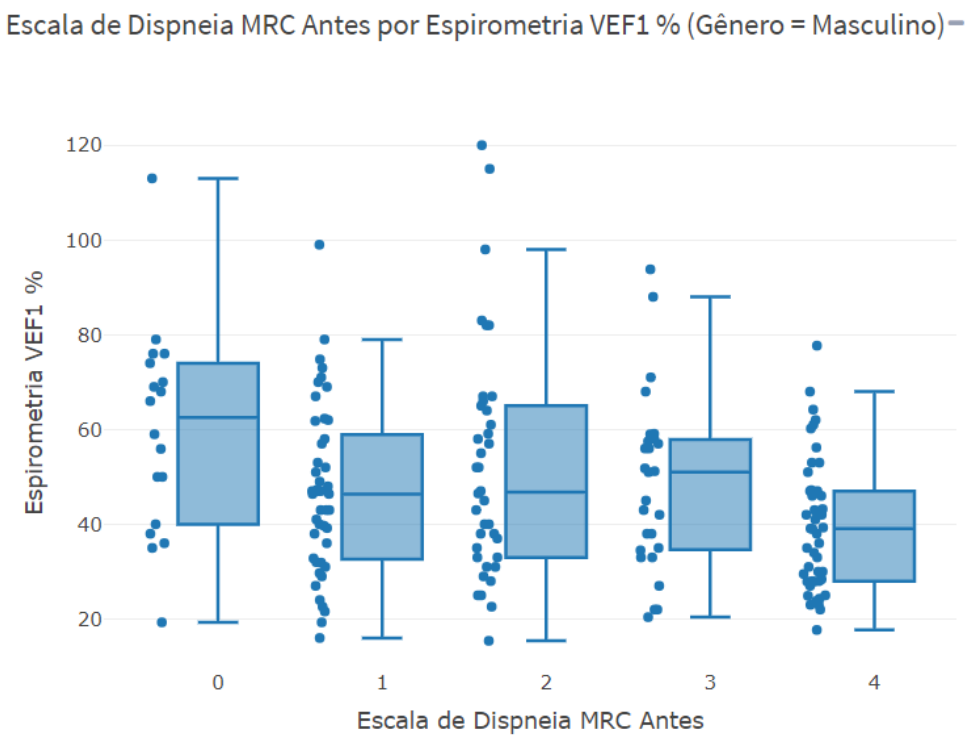
Fonte: elaborado pelo autor

Figura C-5 - Escala de dispneia por espirometria VEF1 % no gênero feminino.



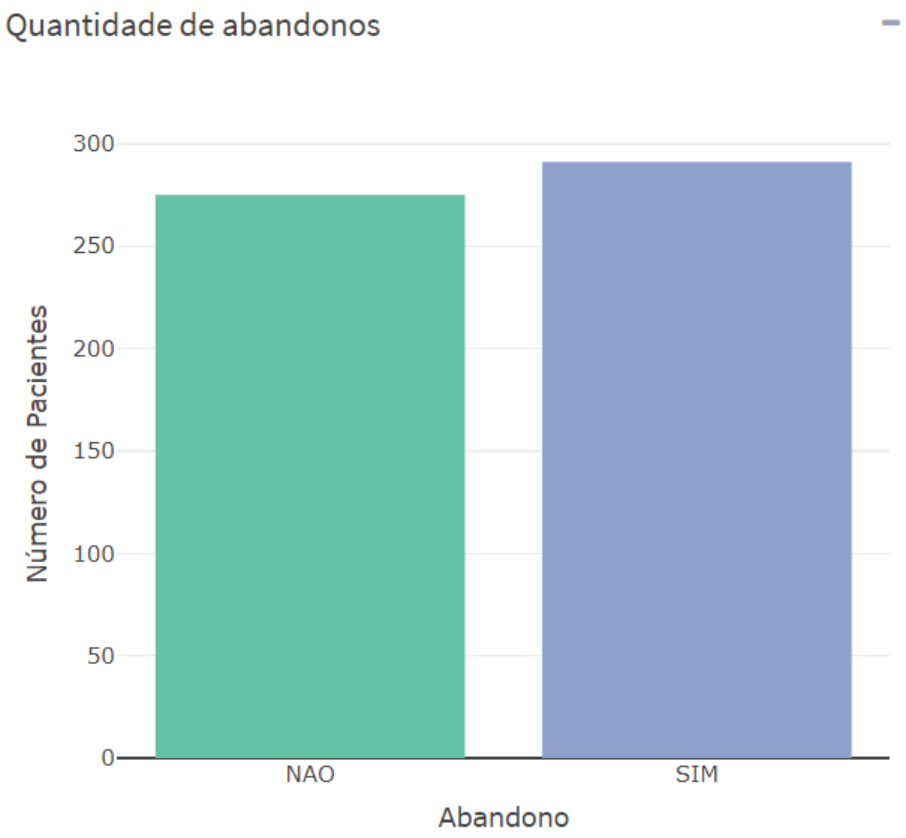
Fonte: elaborado pelo autor

Figura C-6 - Escala de dispneia por espirometria VEF1 % no gênero masculino.



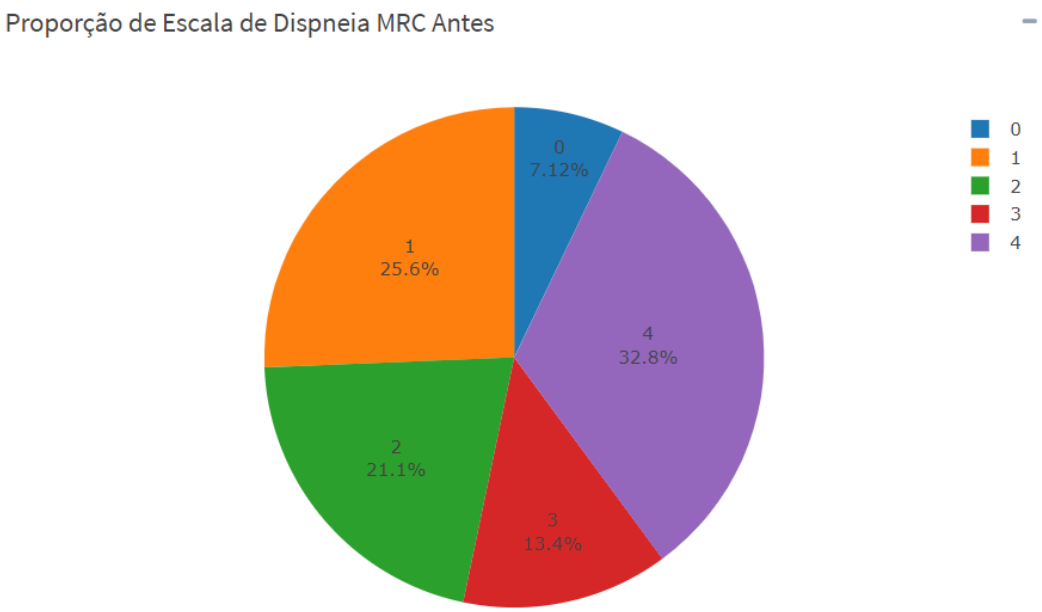
Fonte: elaborado pelo autor

Figura C-7 - Quantidade de abandonos.



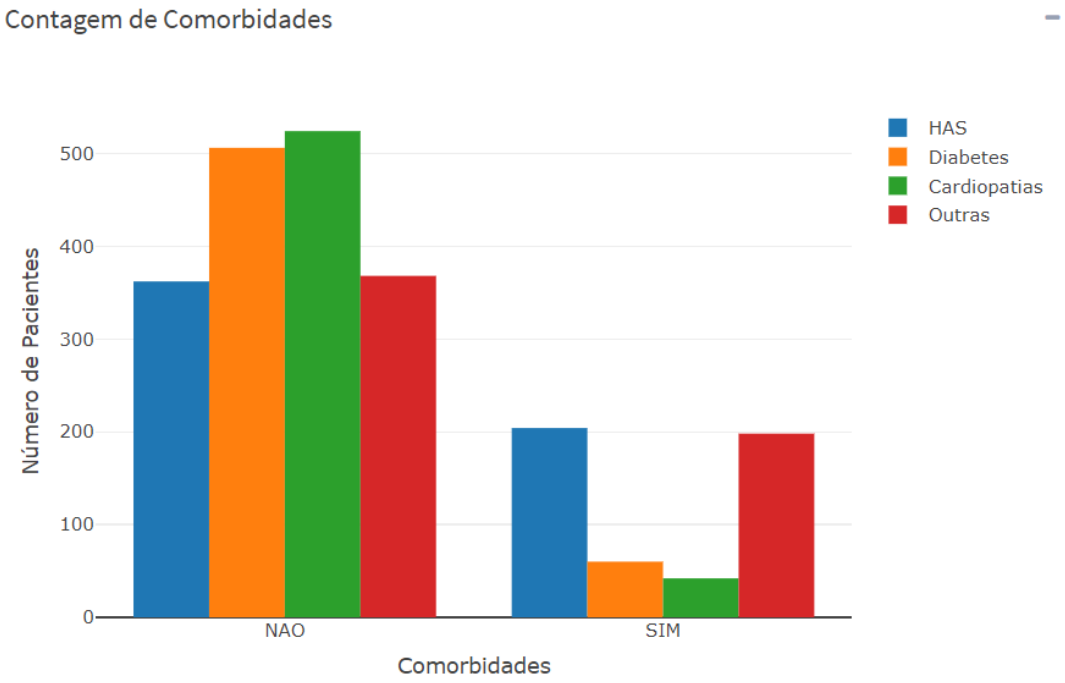
Fonte: elaborado pelo autor

Figura C-8 - Proporção de escala de dispneia.



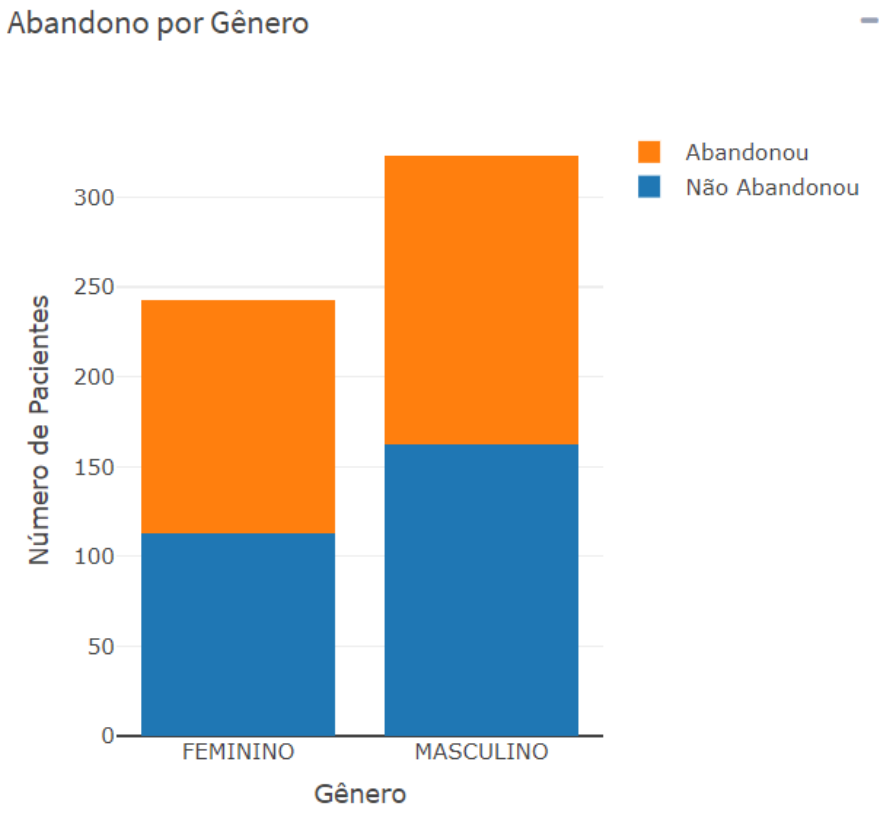
Fonte: elaborado pelo autor

Figura C-9 - Contagem de pacientes por comorbidade.



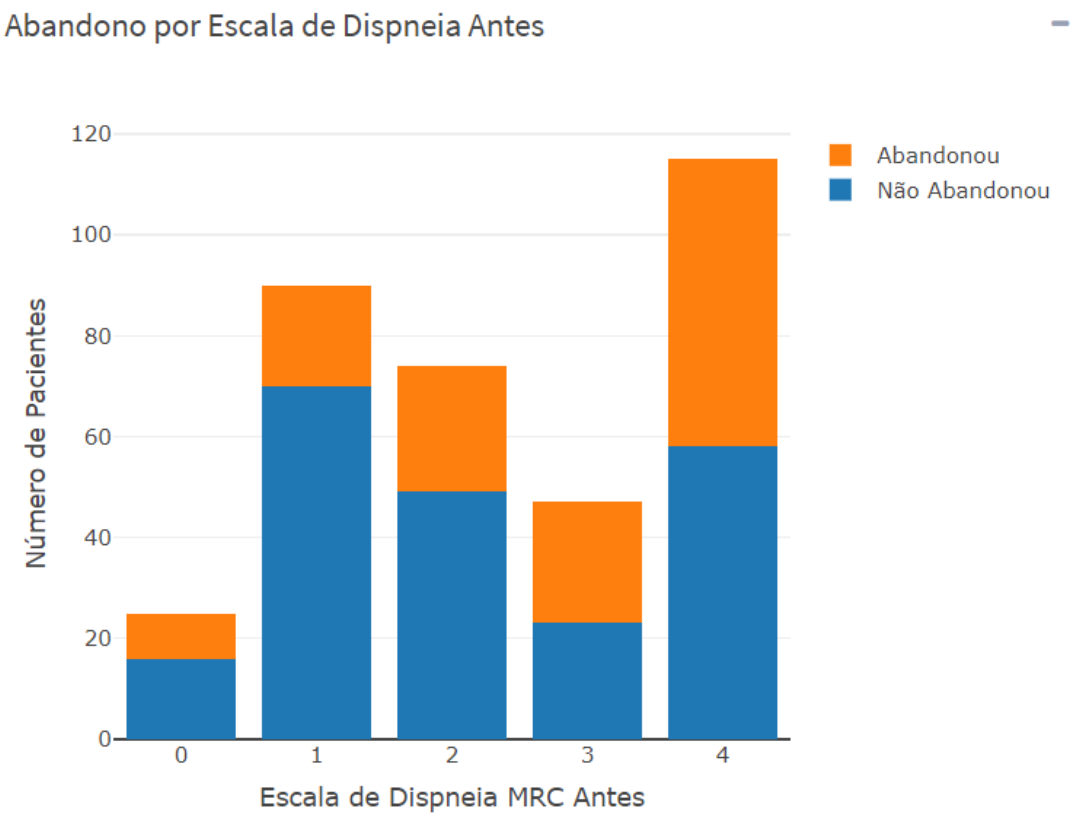
Fonte: elaborado pelo autor

Figura C-10 - Quantidade de abandonos por gênero.



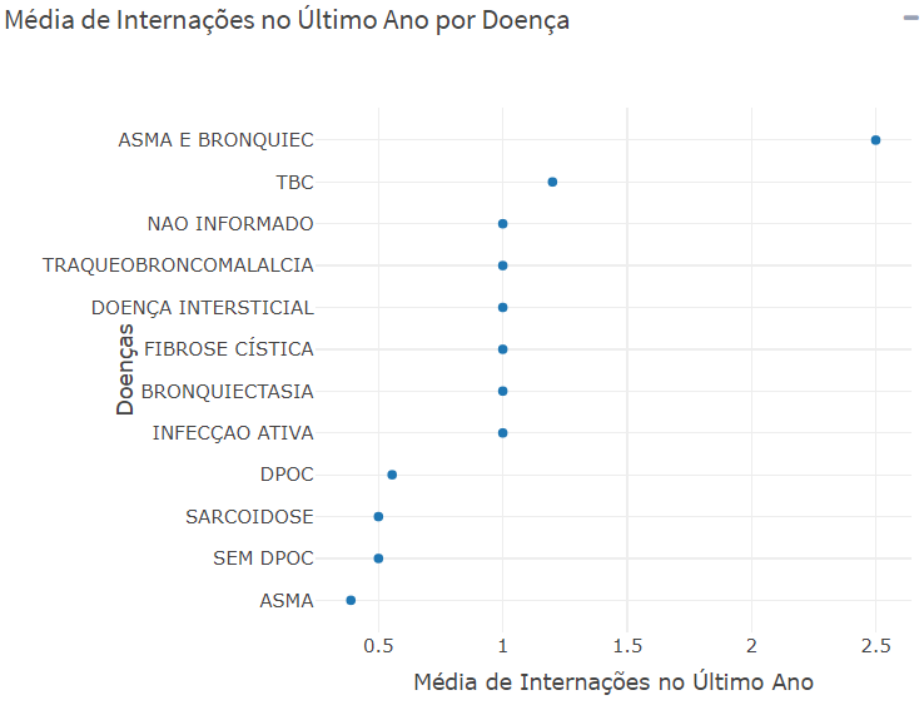
Fonte: elaborado pelo autor

Figura C-11 - Quantidade de abandonos por escala de dispneia.



Fonte: elaborado pelo autor

Figura C-12 - Média de internações no último ano por doença.



Fonte: elaborado pelo autor