PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE

ANA LUÍSA GONÇALVES GOMES COELHO SELEME

IDENTIFICAÇÃO DE PACIENTES DE ALTO CUSTO COM A UTILIZAÇÃO DE DADOS ADMINISTRATIVOS E AUTORREFERIDOS POR MEIO DO APRENDIZADO DE MÁQUINA

CURITIBA
2022

ANA LUÍSA GONÇALVES GOMES COELHO SELEME

IDENTIFICAÇÃO DE PACIENTES DE ALTO CUSTO COM A UTILIZAÇÃO DE DADOS ADMINISTRATIVOS E AUTORREFERIDOS POR MEIO DO APRENDIZADO DE MÁQUINA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Tecnologia em Saúde. Linha de Pesquisa: Informática em Saúde.

Orientadora: Prof.ª Dr.ª Deborah Ribeiro Carvalho

CURITIBA

2022

ANA LUÍSA GONÇALVES GOMES COELHO SELEME

IDENTIFICAÇÃO DE PACIENTES DE ALTO CUSTO COM A UTILIZAÇÃO DE DADOS ADMINISTRATIVOS E AUTORREFERIDOS POR MEIO DO APRENDIZADO DE MÁQUINA

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Tecnologia em Saúde.

COMISSÃO EXAMINADORA

Prof.ª Dr.ª Deborah Ribeiro Carvalho
Orientadora e Presidente - PUCPR

Prof.ª Dr.ª Priscyla Waleska Targino de Azevedo Simões
(Examinador externo)

Prof.ª Dr.ª Rita Cristina Galagarra Berardi
(Examinador externo - suplente)

Prof. Dr. Sérgio Ossamu Ioshii
(Examinador interno)

Prof.ª Dr.ª Claudia Maria Cabral Moro Barra
(Examinador interno - suplente)

Curitiba, 29 de março de 2022.

RESUMO

Introdução: Pacientes com alto custo compreendem entre 1 e 20% da população e são responsáveis por consumir mais da metade dos recursos dos sistemas de saúde. Identificá-los precocemente permite um melhor planejamento financeiro e terapêutico, incluindo ações de caráter preventivo. A utilização de algoritmos com resultados interpretáveis podem auxiliar neste planejamento, trazendo informações sobre os fatores que contribuem para o alto custo. **Objetivo**: Identificar pacientes com alto custo a partir de dados administrativos e autorreferidos por meio do aprendizado de máquina. **Método**: Trata-se de uma pesquisa quantitativa, retrospectiva e de caráter descritivo, conduzida durante a pandemia da covid-19. A população da pesquisa consistiu em 586 pacientes com plano de saúde coletivo empresarial, que responderam um questionário de autoavaliação de saúde. Foram utilizados dados administrativos de pagamento do plano de saúde e autorreferidos em questionário destes pacientes para a construção de um único conjunto de dados, para a aplicação do algoritmo de aprendizado supervisionado *random forest* e utilizado o método gini para avaliar o grau de importâncias das variáveis. A validação dos resultados foi realizada por seis especialistas em gestão de saúde que responderam um questionário sobre a contribuição das principais variáveis para o desfecho alto custo e sobre a importância da utilização de um algoritmo cujos resultados são interpretáveis para a gestão de saúde. **Resultados**: Após pré-processamento das bases, foram utilizados dados de 553 pacientes e construído um conjunto de dados com 63 variáveis com dados de histórico médico pessoal e familiar, saúde mental, sono, hábitos, motivação para mudança e quantidade de procedimentos de saúde realizados no período de janeiro de 2019 a março de 2021, além da apuração daqueles classificados como alto e baixo custo. O algoritmo *random forest* obteve acurácia de 95,47%, com sensibilidade de 95,1% e especificidade de 95,5%. Quando avaliada a contribuição das variáveis, destacaram-se, além do sexo, aquelas relacionadas a saúde mental (depressão, transtorno bipolar, ansiedade e realização de acompanhamento psicológico), uso de bebida alcóolica, sentir dores fortes, desatualização do esquema vacinal, uso de medicamentos e a condição física impedir a realização de exercícios físicos. Os especialistas apresentaram grau de concordância de 0,80 e todos indicaram preferência pela interpretação dos resultados do algoritmo em suas práticas como gestores de saúde. **Conclusão**: O uso de um algoritmo com resultados interpretáveis possibilitou a identificação de variáveis relacionadas ao alto custo, o que permite a utilização destas informações para proposta e implementação de programas de prevenção a doenças e promoção da saúde. Destacaram-se as variáveis relacionadas à saúde mental que, ainda que não sejam a causa do alto custo, estão fortemente relacionadas a este desfecho, evidenciando a importância do debate e incorporação da temática saúde mental dentro das organizações e sistemas de saúde.

Palavras-chave: Inteligência artificial. Predição. Custos de cuidados de saúde. Gestão em Saúde.

ABSTRACT

Introduction: High-cost patients comprise between 1-20% of the population and are responsible for consuming more than half of health systems resources. Their early identification allows better financial and therapeutic planning, including preventive actions. The use of algorithms with interpretable results can help this planning, providing information about the factors that contribute to high cost. **Objective**: Identify high-cost patients using administrative and self-referred data through machine learning. **Methods**: Quantitative, retrospective, and descriptive research, carried out during the covid-19 pandemic. The research population consisted of 586 patients who answered a health assessment questionnaire with a corporate health plan. Administrative data and self-reported data of these patients were used for random forest application, a supervised learning algorithm, and the Gini index was used to assess the importance of the variables. Validation of results was performed by six experts in health management who answered a questionnaire on the contribution of the main variables to being a high cost and on the importance for health management of using an algorithm whose results are interpretable. **Results** The pre-processing results in 553 patients' data for analysis. A dataset with 63 variables was obtained with data on personal and family medical history, mental health, sleep, habits, motivation for change, and the number of health procedures performed in the period, from January 2019 to March 2021, in addition to calculating those classified as high and low cost. The random forest algorithm obtained an accuracy of 95.47%, with a sensitivity of 95.1% and specificity of 95.5%. When evaluating the contribution of variables, in addition to sex, those related to mental health (depression, bipolar disorder, anxiety, and psychological follow-up), use of alcoholic beverages, feeling severe pain, outdated vaccination schedule, use of medications, and physical condition prevent physical exercise. Experts showed a 0.80 agreement degree, and all indicated a preference for interpretation over algorithm results in their practices as health managers. **Conclusion**: The use of an algorithm with interpretable results made it possible to identify variables related to high cost, which allows the use of this information for the proposal and implementation of disease prevention and health promotion programs. The variables related to mental health were highlighted, which, although not the cause of the high cost, are strongly related to this outcome, highlighting the importance of the debate and incorporation of the mental health theme within organizations and health systems.

Keywords: Artificial intelligence. Forecasting. Health care costs. Health Management.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANS - Agência Nacional de Saúde Suplementar
AUC - *Area Under the Curve*
CID - Classificação Internacional de Doenças
CVC - Coeficiente de Validade de Conteúdo
ID - Identificação
IESS - Instituto de Estudos de Saúde Suplementar
IMC - Índice de Massa Corporal
FN - Falso negativo
FP - Falso positivo
GBM - *Gradient boosting machine*
LASSO - *Least Absolute Shrinkage and Selection Operator*
LSLR - *Least square linear regression*
OCDE - Organização para a Cooperação e Desenvolvimento Econômico
OMS - Organização Mundial da Saúde
ROC - Curva Característica de Operação do Receptor
SUS - Sistema Único de Saúde
TCLE - Termo de Consentimento Livre e Esclarecido
VN - Verdadeiro negativo
VP - Verdadeiro positivo
WEKA - *Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Devido à mudança no perfil epidemiológico das populações, em especial com o aumento da prevalência das doenças crônicas, os sistemas de saúde têm enfrentado desafios para oferecer saúde de forma sustentável, com qualidade e focada no paciente (SALISBURY et al., 2011). Diante disso, é importante o olhar para aqueles pacientes que mais demandam por serviços de saúde, conhecidos como pacientes com alto custo, que compreendem entre 1 e 20% dos pacientes atendidos pelos sistemas de saúde, são responsáveis por consumir mais de 50% dos recursos e precisam de intervenções específicas para atender suas demandas e evitar desperdícios (SMEETS et al., 2020).

A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) tem como prioridade as intervenções com foco nestes pacientes, devido ao seu potencial de efetivamente conter o rápido crescimento dos custos com saúde. Portanto, uma abordagem válida, confiável e implementável para prever com precisão quem serão os pacientes com este perfil de alto consumo de recursos é bastante importante para projetar ações sensíveis à redução de custos (TANKE et al., 2019), mantendo a qualidade assistencial.

Muitas vezes estes pacientes recebem um cuidado inapropriado e desnecessário para a severidade de suas doenças, o que comprova a necessidade de um maior conhecimento desta população (SIEKMAN; HILGER, 2018). Identificá-los precocemente por meio de modelos preditivos pode evitar desfechos indesejados e garantir um melhor planejamento terapêutico e financeiro (MOTURU; LIU; JOHNSON, 2008).

Estes modelos dependem majoritariamente de dados administrativos, com base em: diagnósticos, histórico de utilização dos serviços e seus respectivos custos (FERVER; BURTON.; JESILOW, 2009), que têm como objetivo o pagamento a prestadores (CARVALHO; DALAGASSA; SILVA, 2015). No Brasil, a informação sobre diagnóstico, bastante presente nas bases de dados de outros países (FERVER; BURTON; JESILOW, 2009), é limitada, pois não é obrigatório o fornecimento da classificação internacional das doenças em guias ambulatoriais, o que dificulta ainda mais a identificação das características epidemiológicas neste cenário (CARVALHO; DALAGASSA; SILVA, 2015).

Outrossim, as bases de dados administrativos não contemplam dados clínicos, psicossociais ou comportamentais, considerados importantes para identificação de

pacientes com alto custo, pois torna os métodos mais efetivos, ampliando a qualidade das predições, dada a possibilidade destes fatores estarem associados a um maior custo (BATES et al., 2014). Estes dados, apesar de não estarem presentes nas bases de dados administrativos, podem ser coletadas por meio de questionários de autoavaliação de saúde, conhecidos como dados autorreferidos (BOSCARDIN et al., 2015).

Estudos realizados a partir de questionários em populações americanas, demonstram que os dados autorreferidos contribuem para a predição de custos (PERRIN et al., 2011; DeSALVO et al., 2009). Com a evolução digital, estes dados passaram a ser coletados com maior facilidade, por meio de websites ou aplicativos em aparelhos celulares, gerando uma grande quantidade de informações de saúde (WERE; KAMANO; VEDANTHAN, 2016).

Os avanços tecnológicos não se destacam somente na coleta dos dados, mas também nas análises destes, com o advento das técnicas oriundas da inteligência artificial (SUSHIMITA et al., 2015). Dentre elas, estão os algoritmos de aprendizado de máquina de aprendizado supervisionado (MORID et al., 2018). Neste tipo de aprendizado de máquina, o desfecho de um conjunto de dados é conhecido, existindo um valor da variável resposta a ser predito, ou seja, no conjunto de dados estão disponíveis as variáveis preditoras e a variável de interesse, responsável por guiar a análise (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Este tipo de aprendizado se destaca para a predição de pacientes de alto custo (BERTSIMAS et al., 2008) por minimizarem as limitações dos testes estatísticos.

Embora os modelos estatísticos, principalmente os modelos de regressão, tenham êxito em suas predições, apresentam alguns desafios importantes: o primeiro é a capacidade limitada de trabalhar com várias variáveis independentes e suas fortes correlações, o que gera multicolinearidade (CHECHULIN et al., 2014). O segundo, compreende a natureza dos dados de saúde, em que valores diversos estão presentes, tornando sua distribuição assimétrica. É comum os dados de saúde apresentarem valores extremos, com cauda a direita e, apesar de avanços nas técnicas estatísticas para acomodar esta distribuição, este método não é capaz de performar melhor que o aprendizado supervisionado (MORID et al., 2018).

No trabalho conduzido por Chechulin e colaboradores (2014), com o objetivo de predizer pacientes com risco de se tornarem alto custo no Canadá por meio de uma análise de regressão logística, os autores reiteram os pressupostos acima como limitação metodológica: a grande quantidade de variáveis e os numerosos

requisitos relacionados aos dados para execução do modelo, que são facilmente mitigadas pelo emprego de algoritmos de aprendizado de máquina.

Quanto às ações empregadas para o gerenciamento do custo e das condições de saúde de pacientes com alto custo ou, ainda, daqueles que serão alto custo no futuro, é necessário o desenvolvimento de modelos de cuidado com base nas necessidades médicas e socioculturais, bem como suas preferências. Transformar modelos em ganhos clínicos e financeiros exigirá bancos de dados bastante abrangentes com informações de pagamento, clínicas e demais fatores que podem determinar o estado de saúde dos indivíduos (KHULLAR; KAUSHAL, 2018; SMITH et al., 2019).

Propostas de modelos com o enriquecimento do conjunto de dados vêm sendo desenvolvidas com o objetivo de avaliar o impacto de outras variáveis, que não somente aquelas relacionadas a custos, nos modelos de predição de pacientes de alto custo (MOHNEN et al., 2020; LUO et al., 2020; KIM; PARK, 2019). Ainda que as variáveis de custo permaneçam como forte preditoras, a adição de dados autorreferidos pelos pacientes com informações de saúde e estilo de vida ainda é pouco explorada e, por este motivo, este trabalho pretende endereçar esta lacuna.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Identificar pacientes de alto custo a partir de dados administrativos e autorreferidos por meio do aprendizado de máquina.

1.1.2 Objetivos específicos

Os objetivos específicos desta pesquisa são:

- Propor um modelo a partir de um algoritmo de aprendizado de máquina;
- Avaliar a importância das variáveis para a construção do modelo;

1.2 CONTRIBUIÇÃO CIENTÍFICA E SOCIAL

Esta pesquisa contribui cientificamente ao fomentar o uso e novos desenvolvimentos de estratégias de inteligência artificial para suporte às decisões em saúde. Sua principal contribuição social é a possibilidade de melhoria no gerenciamento de saúde da população, permitindo ações de prevenção e não ações reativas, comumente empregadas após a ocorrência do alto custo. A incorporação de dados de questionário pode permitir ações direcionadas de saúde podendo, os gestores, alocar os recursos necessários para atendimento, ao mesmo tempo que inserem estes indivíduos em programas, com linhas de cuidado adequadas as suas necessidades. A possibilidade de relacionar informações sobre prontidão para mudanças de hábitos torna-se relevante, uma vez que a adesão ao tratamento é primordial para o êxito dos programas de saúde.

Adicionalmente, o modelo proposto pode ser reproduzido em outros contextos, seja na esfera pública ou privada, bastando apenas a disponibilização das bases de dados administrativos e de questionário de saúde, com resultados autorreferidos por pacientes. Deste modo, torna-se relevante considerando o atual cenário de financiamento e gerenciamento da saúde em ambas as esferas.

2 REFERENCIAL TEÓRICO

Nesta seção serão abordados os conceitos que sustentam esta dissertação, considerando o contexto de saúde brasileiro e o papel da saúde suplementar, para o entendimento do cenário de saúde no país. Posteriormente, são caracterizados os pacientes com alto custo, para a compreensão de sua complexidade e seu impacto nos sistemas de saúde.

2.1 SAÚDE SUPLEMENTAR

O sistema de saúde no Brasil é caracterizado por duas frentes: o Sistema Único de Saúde (SUS), que tem seu financiamento público realizado por meio de tributos pagos pela população e coletados pelo Estado, e o sistema privado, financiado por meio do pagamento direto ao provedor ou seguro saúde (DUARTE et al., 2017).

A saúde como direito fundamental de todo cidadão foi definida pela Constituição de 1988 e, desde então, o Estado experimenta dificuldades para atingir o volume de recursos necessários à garantia dos serviços ofertados à população (MENDES, 2012). Frente a este fato, o sistema de saúde suplementar desempenha importante papel na prestação dos serviços assistenciais no Brasil, pois absorve a demanda da esfera pública (ZIROLDO et al., 2013).

O sistema de saúde suplementar é responsável pelo atendimento de aproximadamente 24% da população brasileira sendo, destes, 68% beneficiários de planos coletivos empresariais e, portanto, com acesso ao sistema mediante vínculo empregatício (IESS, 2021b). Apesar da instabilidade econômica na qual se encontra o país devido a pandemia do novo coronavírus (CAETANO et al., 2020), o número de beneficiários na saúde suplementar vem aumentando, principalmente nesta modalidade de contratação (IESS, 2021b), o que evidencia que as empresas estão buscando manter e garantir a assistência médica privada para seus colaboradores, ainda que acarrete um maior custo para suas operações.

A sinistralidade, índice calculado pela razão entre o custo da assistência e as receitas com planos de saúde (PIRES, 2008), vem aumentando anualmente e é a base para a revisão financeira dos contratos de planos de saúde (ARAÚJO; SILVA, 2018). Este aumento afeta diretamente as empresas que contratam planos de saúde para os seus funcionários, pois sofrem reajustes em seus contratos e, por este motivo,

têm direcionado uma maior parte de seus orçamentos para gestão de saúde corporativa, com investimentos em programas de saúde dentro das empresas (IESS, 2021a).

A gestão de saúde corporativa abrange os objetivos de melhorar a qualidade de vida dos trabalhadores e conter custos (TAVARES; KAMIMURA, 2014). Inúmeros estudos que envolvem desde a implantação de programas de gerenciamento de casos, até ambulatórios assistenciais dentro de empresas demonstram que tais iniciativas reduzem os custos com saúde, custos previdenciários e com absenteísmo, e melhoram os indicadores de saúde dos trabalhadores (SHERMAN; FABIUS, 2012).

A gestão de saúde corporativa orientada por dados ainda é tema recente. A utilização de dados de saúde oriundos da saúde ocupacional já vêm sendo utilizados para a estratificação de risco e mapeamento epidemiológico (MENDES; TEIXEIRA; BONFATTI, 2017). Os dados administrativos de planos de saúde também são fonte de informação para as áreas de saúde das empresas. Entretanto, com a implementação da Lei Geral de Proteção de Dados, em vigor desde agosto de 2020, operadoras de planos de saúde e empresas passaram a ter maiores dificuldades em tramitar estas informações, uma vez que o tratamento dos dados sensíveis passou a ser regulamentado.

Ademais, devido ao prazo de contestação das contas médicas, por meio das glosas e recursos de glosas (RODRIGUES et al., 2018), os dados administrativos chegam até as empresas com até 60 dias da realização dos eventos pelos pacientes, o que faz com que a gestão e intervenções de saúde baseadas nestas informações ocorra tardiamente.

O fato é que a despesa com saúde no mundo permanecerá crescendo, agravada pelo envelhecimento da população e suas doenças complexas relacionadas. A gestão da saúde da população, incluindo análises de tendências, qualidade e custos, são importantes para que os tomadores de decisão consigam melhorar os processos e a coordenação do atendimento diante deste cenário (SHAHID; RAPPON; BERTA, 2019), seja ele no âmbito público ou privado. Em um país em que há um sistema de saúde universal, é do interesse de todos que a saúde corporativa e a saúde suplementar sejam exitosas em suas gestões, para não ocorrer a sobrecarga no sistema de saúde público, conforme exposto por Ziroldo e colaboradores (2013).

2.2 PACIENTES COM ALTO CUSTO

Focar em pacientes com maiores necessidades de atendimento e com alto custo assistencial tem sido uma das estratégias para a redução dos custos com saúde (FIGUEROA; ZHOU; JHA, 2019), uma vez que são responsáveis pela utilização desproporcional de recursos e têm necessidades significativas de atendimento (BLUMENTHAL et al., 2016). A definição de pacientes com alto custo sofre variações, sendo considerados aqueles que representam 1%, 5%, 10% ou 20% do total dos custos (WAMMES et al., 2017).

Quanto à sua nomenclatura, não existe um consenso na literatura, podendo os pacientes com esta característica serem definidos pelas seguintes terminologias: superutilizador, superusuário, beneficiário ou paciente de alto custo, paciente de alto custo e alta necessidade (LEE et al., 2018). Neste trabalho, será utilizado o termo paciente de alto custo para caracterizar os pacientes com os atributos expostos acima.

Em um estudo conduzido por Siekman e Hilger (2018), os autores constataram que apenas 1% dos maiores utilizadores do sistema de saúde americano consome aproximadamente 23% dos recursos, gastando dez vezes mais do que os demais usuários. Ressaltam, ainda, que estes pacientes custosos muitas vezes recebem um cuidado inadequado e/ou desnecessário.

Os pacientes considerados de alto custo são caracterizados como uma população com inúmeras doenças crônicas e prevalência de doenças mentais (WAMMES et al., 2017). Entretanto, estudos demonstram que estes pacientes têm características bastante heterogêneas, com atributos biopsicossociais distintos e não podem ser estereotipados e classificados somente como pacientes idosos e/ou com multimorbidades (BLUMENTHAL et al., 2016; SMEETS et al., 2020).

No Brasil, em estudo publicado pelo Instituto de Estudos de Saúde Suplementar (IESS, 2017) com uma operadora de plano de saúde de grande porte, constatou-se que 66,5% dos gastos assistenciais eram consumidos pelo atendimento de apenas 5% de seus beneficiários. Os principais fatores relacionados ao alto consumo destes recursos eram: envelhecimento, doenças crônicas e a frequência de internação de pacientes com multimorbidades.

Em estudo que analisou os pacientes mais custosos dos sistemas de saúde de países desenvolvidos a fim de identificar um perfil, constatou-se que estes pacientes são, em sua maioria, do sexo feminino, com multimorbidades e baixo

status socioeconômico. As doenças predominantes neste grupo eram: doenças cardiovasculares, musculoesqueléticas e neoplasias. Não foi identificado um padrão em relação à concentração dos custos destes pacientes dentre os países analisados. No Japão, por exemplo, 5% dos maiores custos eram responsáveis pelo consumo de 41% dos recursos, contra 60% no Canadá. Estas diferenças estão relacionadas a variação dos custos dos pacientes que compreendem o 1% mais custoso (TANKE et al., 2019).

A multimorbidade é um fator representativo nos estudos sobre pacientes com alto custo. Em análise dos dados de pacientes cobertos pelo sistema de saúde dos veteranos de guerra, nos Estados Unidos, pesquisadores identificaram um custo incremental de aproximadamente US$1.774 no total dos custos para cada sistema do corpo humano afetado por uma doença crônica. Ainda, para cada doença adicional, foram constatados aumento médio de 0,2 hospitalizações, 0,4 visitas ao pronto-socorro, 1,5 consultas ao médico de atenção primária e 0,7 consultas especializadas (ZULMAN et al., 2015).

Outro destaque importante deve ser dado aos transtornos mentais, que são observados nestes pacientes, associados ou não a outras doenças crônicas (JOYNT et al., 2013; WAMMES et al., 2018). Em um estudo que analisou dados de pacientes beneficiários dos sistemas Medicare e Medicaid dos Estados Unidos, concluiu que estes pacientes têm motivadores diferentes para a utilização dos recursos com saúde. Os gastos com saúde no Medicare, que assegura assistência médica para idosos, são impulsionados em grande parte por atendimento a sequelas de doenças e múltiplas condições crônicas, enquanto no Medicaid, focado em uma população mais jovem e vulnerável, os transtornos mentais predominam (KHULLAR; KAUSHAL, 2018).

Além do estabelecimento de um perfil, é importante entender por que estes pacientes permanecem gastando, em contrapartida aos que têm custos transitórios, para que se possa definir quais intervenções serão úteis e em que momento devem ser feitas. São considerados pacientes com custos persistentes aqueles que gastam por mais de um ano (WAMMES et al., 2017). Entre 20 e 60% dos pacientes com alto custo permanecem como alto custo no ano seguinte (COUGHLIN; LONG, 2009; FIGUEROA; ZHOU; JHA, 2019).

Pacientes com câncer, por exemplo, apesar de estarem dentre os mais custosos dos sistemas de saúde, possuem padrões de gastos diferentes dos

demais. Tendem a fazer picos de custo no início do diagnóstico e voltam a gastar perto de sua morte (DE OLIVEIRA et al., 2017). Wodchis e colaboradores (2016) também analisaram a trajetória de custo de pacientes com câncer e demonstraram que comumente são considerados baixo custo antes e após o tratamento oncológico, evidenciando a necessidade de análises separadas em análises de tendências e predição de custo.

Quanto às intervenções necessárias para estes pacientes, os programas de saúde precisam considerar os desafios dos indivíduos perante suas condições (ZULMAN et al., 2015) para que sejam efetivos e não acarretem num desperdício de recursos ainda maior. Intervenções e diretrizes para o gerenciamento de múltiplas condições crônicas, característica bastante evidente nos pacientes com alto custo, enfatizam a importância de considerar as preferências do indivíduo, suas dificuldades funcionais e seu prognóstico ao desenvolver planos de cuidados (SMITH et al., 2019).

3 REFERENCIAL METODOLÓGICO

Neste capítulo serão abordados os conceitos acerca das bases de dados disponíveis para construção de modelos de identificação de pacientes com alto custo, bem como referências sobre *machine learning* e o algoritmo proposto para esta pesquisa.

3.1 BASES DE DADOS

As bases de dados comumente utilizadas para predição em saúde envolvem o uso de dados administrativos (HIBBARD et al., 2016), que compreendem bancos de dados de saúde nacionais, a exemplo do Medicare e Medicaid, nos Estados Unidos, bem como registros de seguros privados, que incluem os tipos e as quantidades de serviços utilizados, como consultas, hospitalizações e exames realizados, as cobranças realizadas, em quais prestadores de serviço e até mesmos problemas médicos tratados (DIEHR et al., 1999).

Seu uso é benéfico, pois classifica, de forma rápida, os pacientes de maior risco e que precisam de avaliação adicional (KIM et al., 2017). Entretanto, é limitado, pois utiliza-se somente do passado para fazer projeções futuras (HIBBARD et al., 2016). Ainda, têm diversas características que os tornam um desafio para serem analisados, considerando que as pessoas entram e saem do sistema de acordo com a sua elegibilidade, o que não permite uma avaliação temporal dos seus dados (DIEHR et al., 1999).

Cerca de 13% dos indivíduos cobertos por seguro saúde nos Estados Unidos mudam seu plano de saúde em um ano e aproximadamente 9% mudam seu prestador de serviços de saúde (CUNNINGHAM, 2017). No Brasil, considerando que 67% daqueles assistidos são elegíveis ao plano de saúde por meio de seus empregadores (ANS, 2019), seus dados administrativos não estão disponíveis para os novos fornecedores e médicos quando assumem os cuidados destes pacientes.

Ademais, os dados administrativos têm como finalidade o controle administrativo, visando ao pagamento de prestadores e à gestão de contratos e custos. No Brasil, uma decisão do ano de 2007 do Conselho Federal de Medicina desobrigou a notificação do código da classificação internacional de doenças (CID) em guias ambulatoriais e, por este motivo, tornou ainda mais difícil a identificação de dados

epidemiológicos em bases de dados administrativos (CARVALHO; DALAGASSA; SILVA, 2015).

Os custos com saúde podem refletir como está o estado de saúde de um indivíduo ou população. Porém, a compreensão de quais fatores contribuem para aumentos nestes custos pode fornecer uma visão sobre os fatores de risco e potenciais pontos de partida para medidas preventivas (LENTZ et al., 2019).

Por estes motivos, estudos têm utilizado questionários de saúde com dados autorreferidos por pacientes. Relatos da aplicação de questionários em populações americanas, demonstram que este dados apoiam a predição de custos em saúde de maneira bastante eficiente. (BOSCARDIN et al., 2015; PERRIN et al., 2011; DeSALVO et al., 2009).

De acordo com Bates et al. (2014), para implementar métodos efetivos de identificação de pacientes com alto custo, alguns pontos devem ser considerados, como fatores comportamentais e socioeconômicos, que podem estar associados a uma maior chance de alto custo e mudam significativamente a qualidade das predições. Tais informações podem ser coletadas em questionário de autoavaliação.

Em um trabalho com objetivo de identificar fatores associados aos custos persistentes de pacientes com dor musculoesquelética, pesquisadores adicionaram ao conjunto de dados administrativos, resultados de questionário de saúde com avaliação da saúde física e mental, criando variáveis divididas em fatores de risco não modificáveis e modificáveis, uma vez que estes podem ser passíveis de mudanças e intervenções de saúde. Concluíram que modelos que identificam prospectivamente estes fatores pode melhorar a gestão dos custos e o cuidado dos pacientes propensos ao alto custo (LENTZ et al., 2019).

Em outro trabalho, conduzido por Peter Cunningham (2017) para identificar o quanto os dados autorreferidos sobre saúde e comportamento de pacientes de um seguro privado podem prever se eles incorrerão em alto custo no ano seguinte, demonstrou que, na ausência de dados sobre uso e despesas anteriores, as medidas do estado de saúde relatadas pelo paciente eram boas preditoras de custos futuros.

Em um estudo conduzido no Japão (OSAWA et al., 2020) os pesquisadores identificaram que a incorporação de dados clínicos, demográficos e de questionários de saúde, contendo informações sobre tabagismo, uso de medicações e histórico médico, teve uma performance superior ao uso de dados administrativos isoladamente em modelos preditores de custos futuros. A mesma hipótese foi testada por Kim e Park

(2019), com a utilização de dados de check-up médico, exames laboratoriais e dados autorreferidos.

As variáveis relacionadas a custo envolvem, comumente: custo total, custo com medicações, custo com internações e custos com atendimento ambulatorial, além de custos fracionados em períodos, como: últimos três ou seis meses. Já as variáveis não relacionadas a custo compreendem: idade, sexo, grupo diagnóstico, quantidade de consultas e internações, índices de comorbidade e dados autorreferidos sobre saúde física e mental (MORID et al., 2018).

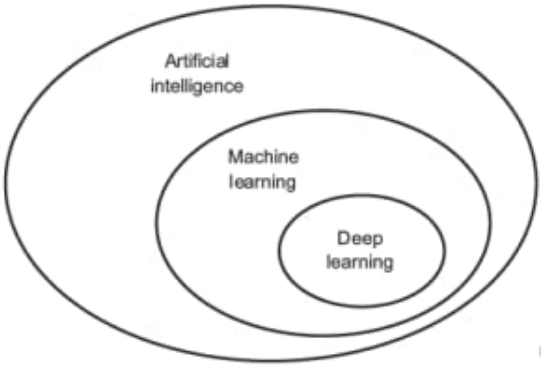
Embora a performance dos modelos preditivos possa variar de acordo com as bases de dados utilizadas, as informações relacionadas as doenças são consideradas preditores chave para identificação de pacientes com alto custo (MOTURU; LIU; JOHNSON, 2008).

Os estudos incluem uma grande variedade de variáveis para composição dos modelos (MORID et al., 2018). É importante destacar, entretanto, que uma grande quantidade de variáveis não implica em melhores modelos. Bertsimas e colaboradores (2008) avaliaram inúmeras possibilidades e concluíram que a performance de 21 variáveis era similar a performance de um modelo que utilizou 1542.

3.2 *MACHINE LEARNING*

Aprendizado de máquina, *ou machine learning*, foi definido inicialmente por Arthur Samuel (1959) como um “campo de estudo que dá ao computador a habilidade de aprender sem ser explicitamente programado”, e é apresentado por Cholet (2018) como uma subárea da inteligência artificial, conforme representado na Figura 1.

Figura 1 - Inteligência artificial e representação do *Machine Learning*



Fonte: Chollet (2018).

Algoritmos de *machine learning* apresentam potencial para identificar relações complexas e não lineares presentes nos dados existentes na área da saúde, com consequências positivas na performance preditiva dos modelos (SANTOS, 2018).

Dentre as formas de aprendizado que orientam os algoritmos de *machine learning* destacam-se o aprendizado supervisionado e não supervisionado. O aprendizado supervisionado caracteriza-se por conter no conjunto de dados as respostas de interesse para guiar as análises (HASTIE; TIBSHIRANI; FRIEDMAN, 2008), ou seja, contém as respostas observadas, rotuladas, sendo o objetivo do algoritmo aprender a partir destes exemplos e desenvolver habilidade de responder corretamente a partir de novos dados de entrada (MARSLAND, 2015). Já no aprendizado não supervisionado, as respostas de interesse não estão presentes, fazendo com que o algoritmo reconheça padrões a partir de um conjunto de dados não rotulados (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

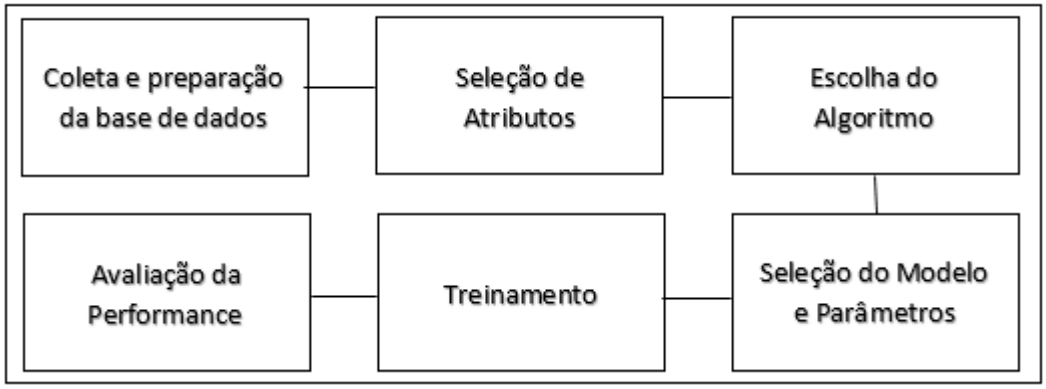
O aprendizado não supervisionado vem sendo frequentemente utilizado na área da saúde para problemas de agrupamento, como nos estudos de identificação de subgrupos de pacientes com doença cardíaca em uso de telemedicina (BOSE; RADHAKRISHNAN, 2018) ou características de equipes médicas com síndrome de *Burnout,* fadiga intensa relacionada ao trabalho (LEE et al., 2016).

Já o aprendizado supervisionado tem se destacado como metodologia para predição de doenças, mortalidade e pacientes de alto custo (SANTOS, 2018; MORID et al., 2018). Cada método de treinamento tem uma finalidade e, no caso do aprendizado supervisionado, duas atividades são possíveis: classificação, para uma variável de saída discreta, e regressão, para variáveis de saída contínuas (ALZUBI; NAYYAR; KUMAR, 2018). Neste projeto, será abordada a tarefa de classificação

binária, que tem por objetivo decidir em qual classe uma nova observação pertence considerando duas classes possíveis (SOKOLOVA; LAPALME, 2009): paciente com alto custo e baixo custo.

Independentemente do tipo de aprendizado utilizado, um modelo de *machine learning* consiste em seis passos, conforme detalhado na Figura 2 e explicação subsequente.

Figura 2 - Passos de um modelo genérico de *Machine Learning*



Fonte: Adaptado de Alzubi; Nayyar; Kumar (2018).

A coleta e preparação da base de dados envolve a análise dos dados disponíveis e preparação para que possam ser fornecidos como entrada para o algoritmo (ALZUBI; NAYYAR; KUMAR, 2018). Após esta preparação, são selecionados os atributos, considerando tanto aspectos quantitativos, quanto qualitativos. O objetivo aqui, é selecionar variáveis que estejam relacionadas com o desfecho a eliminar aquelas redundantes ou irrelevantes (PANAY et al., 2020).

Já a escolha do algoritmo depende do problema a ser resolvido. Na área da saúde, muitos relacionamentos de interesse são lineares, ou seja, com relações diretas, como a relação entre o índice de massa corpórea e o risco de diabetes, podendo exigir o emprego de modelos relativamente simples, com resultados de fácil interpretação. Em modelos mais complexos, como identificação de imagens, identificar a relação entre as variáveis e o desfecho podem deixar de ser relevantes.

Os resultados alcançados pelos algoritmos podem ser classificados como facilmente interpretáveis, conhecidos como algoritmos auditáveis, e aqueles de difícil interpretação, não auditáveis, conhecidos como algoritmos caixa-preta. A opção por um ou outro algoritmo depende do problema a ser resolvido, da complexidade dos dados e da necessidade de compreensão e explicação de seus resultados (SIDEY-

GIBBONS; SIDEY-GIBBONS, 2019). Os algoritmos auditáveis consomem menos recursos computacionais, ao contrário daqueles caixa-preta, que podem demandar dias para processar e construir modelos, e possuem relações lineares entre os preditores e o desfecho. Os algoritmos caixa-preta são considerados melhores dados complexos, entendidos por aqueles dados com relações não-lineares entre preditores e desfecho, como por exemplo, pixels em imagens ou movimentos capturados por *smartphones* (BEAM; KOHANE 2018; SIDEY-GIBBONS & SIDEY-GIBBONS, 2019).

Após a escolha do algoritmo, são definidos os seus parâmetros, que controlam a complexidade dos modelos, ou seja, o equilíbrio entre viés e variância (SANTOS, 2018). Não há uma fórmula específica para o cálculo destes parâmetros que, dependendo do conjunto de dados, demanda ajustes personalizados para cada situação (ANDRADE, 2013). Um exemplo de definição de parâmetro é a quantidade de árvores de decisão, naquelas baseado em árvores (SANTOS, 2018). Após a configuração dos parâmetros, a fim de minimizar erros, segue-se para a etapa de treinamento (BERGSTRA; BENGIO, 2012).

A etapa de treinamento consiste na separação de parte da base de dados para treinar o algoritmo (ALZUBI; NAYYAR; KUMAR, 2018). Na presença de um grande conjunto de dados, é possível dividir aleatoriamente o conjunto de dados em treinamento, validação e teste. É importante garantir que cada subconjunto seja representativo do total do conjunto de dados, evitando erros de aprendizado e posterior validação. Em conjunto de dados menores, que impossibilitam esta divisão em três partes, são utilizadas técnicas de reamostragem, com o objetivo de aproximar o conjunto de validação com a reutilização de observações do conjunto de treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Dentre as técnicas de reamostragem mais frequentes, está a validação cruzada, que divide o conjunto de dados não testados em K partições, com o treinamento ocorrendo K vezes, sendo todas as partições validadas (DOUPE; FAGHMOUS; BASU, 2019). Não há uma recomendação para a quantidade de partições, embora a literatura comumente relate a divisão entre 5 e 10. À medida que este número aumenta, a diferença do tamanho do conjunto de treinamento original e subconjuntos reamostrados se torna menor e, quando esta diferença diminui, o viés da técnica também é menor. Entretanto, quanto maior este número, maior o tempo necessário para a obtenção do resultado (KUHN; JOHSON, 2013).

Por fim, parte-se para o teste de avaliação da performance do modelo com métricas de desempenho como: acurácia, sensibilidade, especificidade e área abaixo

da curva ROC (AUC) (ALZUBI; NAYYAR; KUMAR, 2018; SOKOLOVA; LAPALME, 2009). As métricas para classificação binária, fórmulas de cálculo e seus principais focos de avaliação encontram-se no Quadro 1. Para descrever as métricas, são utilizados os termos verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN). A exatidão de uma classificação é avaliada calculando o número de exemplos de classes reconhecidos corretamente (VP), o número de exemplos reconhecidos corretamente que não pertencem a classe (VN) e exemplos que foram atribuídos incorretamente a classe (FP) ou que não foram reconhecidos como exemplos de classe (FN). Estas contagens formam uma matriz confusão, conforme exemplo do Quadro 2.

Quadro 1 - Métricas para classificações binárias

Métrica	Foco de Avaliação	Fórmula
Acurácia	Concordância entre as classes preditas e observadas	$\frac{VP+VN}{(VP+FN+FP+VN)}$
Sensibilidade	Proporção de verdadeiros positivos dentre todos os indivíduos cuja resposta de interesse foi observada	$\frac{VP}{VP+FN}$
Especificidade	Proporção de verdadeiros negativos dentre todos os indivíduos cuja resposta de interesse (observada) foi ausente	$\frac{VN}{VN+FP}$
AUC	Habilidade do algoritmo de evitar uma classificação errada	$\frac{1}{2}\left(\frac{VP}{VP+FN}\right)+\left(\frac{VN}{VN+FP}\right)$

Fonte: Adaptado de Sokolova e Lapalme (2009).

Quadro 2 - Matriz confusão

Resultado/Classe	Classificado como positivo	Classificado como negativo
Positivo	VP	FP
Negativo	FN	VN

Fonte: Adaptado de Sokolova e Lapalme (2009).

A matriz confusão representa que, se determinada condição está presente no paciente e a sua classificação indica sua presença, o resultado é considerado verdadeiro positivo (VP). Se o paciente não apresenta determinada condição e sua classificação indica sua ausência, é considerado verdadeiro negativo (VN). Por outro lado, se a sua condição está presente e a sua classificação indica sua ausência, é

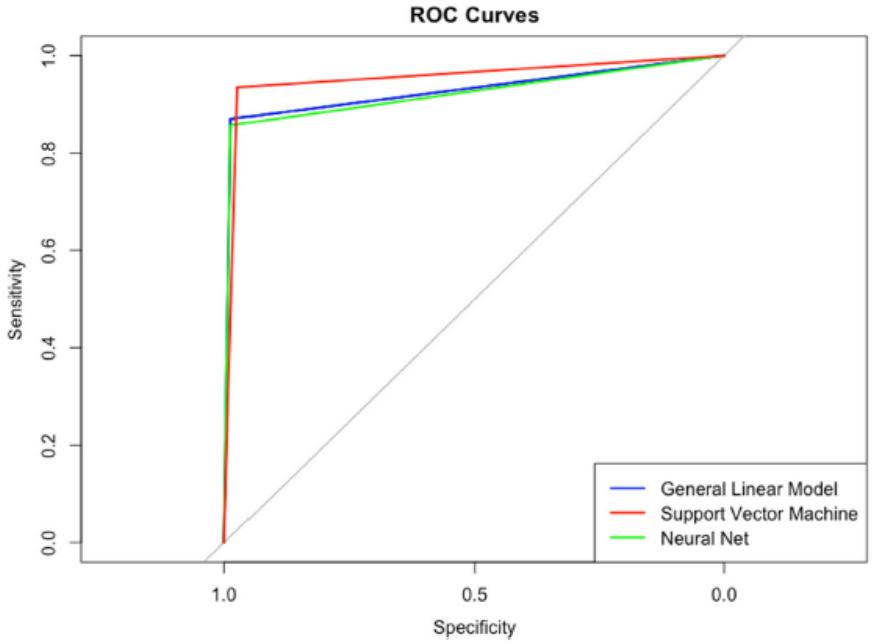
considerado um falso negativo (FN), bem como se sua condição está ausente e sua classificação presente, sendo considerado um falso positivo (FP).

A curva ROC representa a união dos pontos entre sensibilidade e especificidade em um plano cartesiano. No eixo Y encontra-se a sensibilidade e no eixo X, 1 menos a especificidade (1-E). Para cada ponto de corte utilizado pelo teste são calculadas a sensibilidade e a especificidade e inserido um ponto no gráfico que, uma vez unidos, formam a curva ROC (LOPES et al., 2014).

A interpretação da curva ROC é facilitada pelo cálculo da área sob a curva ROC (AUC), que dá um único valor que explica a probabilidade de uma amostra aleatória ser classificada corretamente por um algoritmo: quanto mais próxima de 1 (um), melhor sua performance, ou ainda, quanto mais próxima do canto superior esquerdo, melhor seu poder discriminatório (LOPES et al., 2014; SOKOLOVA; LAPALME, 2009).

Um exemplo de curva ROC encontra-se na Figura 3, demonstrando as combinações entre sensibilidade e especificidade para diferentes algoritmos de *machine learning.* É possível perceber boa performance para todos os algoritmos, com destaque para o algoritmo *Support Vector Machine*.

Figura 3 - Exemplo de curva ROC



Fonte: Sidey-Gibbons e Sidey-Gibbons (2019).

Na próxima seção serão abordados os algoritmos utilizados para modelos de identificação precoce de pacientes com alto custo encontrados na literatura, bem como o algoritmo de escolha para o desenvolvimento dessa pesquisa.

3.3 ALGORITMOS PARA PREDIÇÃO DE PACIENTE COM ALTO CUSTO

Para a tarefa de classificação, como no caso desta dissertação, ao menos uma heurística deve ser utilizada, com base no problema a ser resolvido e nas características dos dados utilizados norteando, assim, a escolha do algoritmo. Diversos algoritmos de aprendizado supervisionado são utilizados para predição de custo em saúde. De modo geral, os algoritmos podem ser agrupados nas seguintes categorias: lineares, como o *Least Absolute Shrinkage and Selection Operator* (LASSO), não lineares, como as redes neurais, e baseados em árvores de decisão, como *gradient boosting* e *random forest* (SANTOS, 2018). Dentre os algoritmos de melhor performance para predição de pacientes com alto custo, encontra-se as redes neurais artificiais (YANG et al., 2018; SHAHID; RAPPON; BERTA, 2019; MORID et al., 2018).

As redes neurais artificiais vêm sendo amplamente exploradas na área da saúde, tanto em estudos epidemiológicos como de diagnóstico e prognóstico médico (SANTOS et al., 2005).

Com seu funcionamento baseado no cérebro humano, as redes neurais são divididas em três camadas de neurônios: uma de entrada, responsável por receber as informações, uma oculta, responsável por extrair padrões e uma de saída, que produz e apresenta os resultados da rede (SHAHID; RAPPON; BERTA, 2019). São modelos de processamento em série distribuídos paralelamente, que exploram diversas hipóteses por meio de regras de aprendizagem que adquirem poder de generalização para reconhecimento de padrões e predição de cenários (HAYKIN, 2001).

O uso das redes neurais envolve grande quantidade de dados, com potencial para correção de dados imprecisos, sendo bastante eficaz em tarefas em que o conjunto de regras não é facilmente definido (ROSAS; BEZERRA; DUARTE-NETO, 2013). Entretanto, é um algoritmo que consome muitos recursos computacionais, o que pode torná-lo uma desvantagem quando comparado a outros algoritmos (THEOBALD, 2017).

Outro fator importante é que, para modelos preditivos em saúde, a interpretabilidade do método, isto é, a compreensão de como as variáveis se relacionamento com o desfecho (YANG et al., 2018; SIDEY-GIBBONS; SIDEY-GIBBONS, 2019), pode ser tão importante quanto sua exatidão. Ainda, a interpretabilidade possibilita um ganho maior de confiança dos usuários finais. Por estes motivos, alguns modelos são construídos com algoritmos auditáveis, mais simples e transparentes, ainda que com menos precisão (PANAY et al., 2019).

Em uma revisão sistemática com avaliação empírica conduzida por Morid e colaboradores (2018) para avaliar diferentes abordagens para predição de custo, os autores concluíram que as redes neurais artificiais, têm uma melhor performance para identificar pacientes com alto custo, seguido pelo algoritmo *grandient boosting*. Ambos os algoritmos são considerados algoritmos caixa-preta, que sacrificam a interpretação dos resultados (PANAY et al., 2020). Dentre outros algoritmos empregados e com bons resultados estão aqueles de regressão linear e de árvores de decisão, como o *random forest.*

Em estudo sobre abordagens de aprendizado supervisionado para predizer pacientes com alto custo e avaliar a consistência temporal destes custos, Yang e colaboradores (2018) utilizaram dados administrativos de pacientes com asma, diabetes, doença pulmonar obstrutiva crônica e hipertensão como entrada para os algoritmos de regressão linear (LR – *least square linear regression* e LASSO – *regularized regression*), *gradient boosting machine* (GBM) e uma rede neural recorrente.

A amostra compreendia dados de 4 anos de 1.734.896 milhões de pacientes do programa de saúde americano Medicaid do estado do Texas, e a conclusão é de que há uma forte correlação temporal nos gastos de pacientes com alto custo, demonstrando a consistência nos gastos destes pacientes. Quanto à performance dos algoritmos utilizados, os autores reforçam que a escolha depende se o objetivo é prever melhor as despesas ou compreender melhor a contribuição das variáveis e reiteram a limitação das redes neurais para este objetivo. Acrescentam, também, que para ações efetivas de saúde, se faz necessária a incorporação de dados no modelo preditivo que melhor ajudem na compreensão das características destes pacientes.

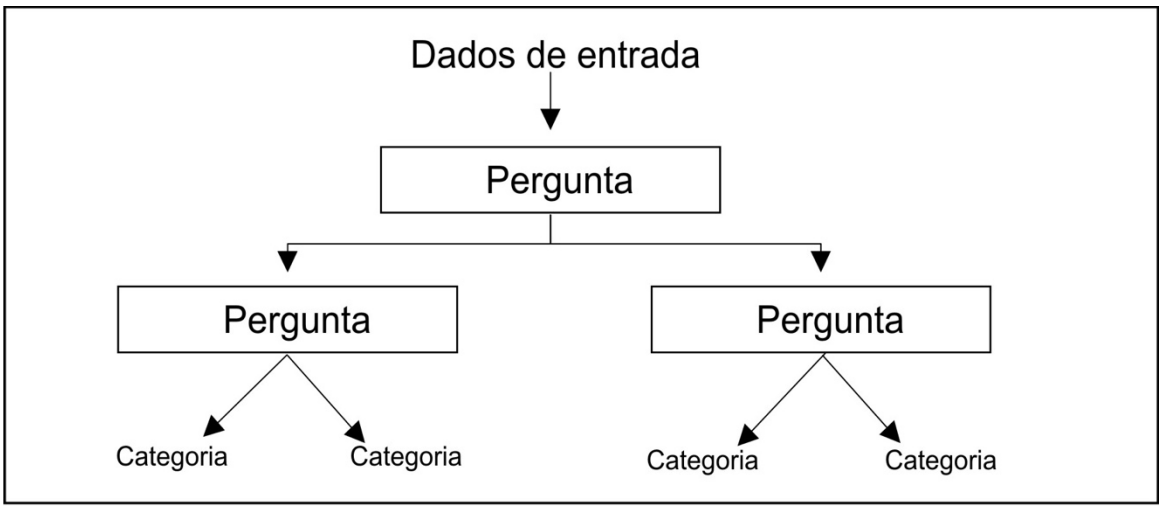
Considerando que os resultados obtidos por algoritmos de *machine learning* podem ser interpretáveis ou não (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019), e, que a literatura relata boa performance dentre uma quantidade razoável de algoritmos, para este trabalho será utilizado o algoritmo *random forest*, pela performance evidente

na literatura, e por ser um algoritmo cujos resultados podem ser interpretados, a fim de avaliar a contribuição das variáveis propostas para o modelo de identificação de pacientes com alto custo.

3.3.1 *Random forest*

O *random forest foi* proposto por Breiman (2001) e é um algoritmo baseado em árvores de decisão. As árvores de decisão são estruturadas em formato de fluxogramas que permitem classificar dados de entrada ou predizer dados de saída considerando os dados de entrada. Os parâmetros aprendidos são os nós, que compreendem perguntas sobre os dados e a resposta é a categoria em que o exemplo se enquadra (CHOLLET, 2018), conforme ilustrado na Figura 4. Em uma árvore de decisão, cada nó é dividido usando a melhor divisão entre todas as variáveis.

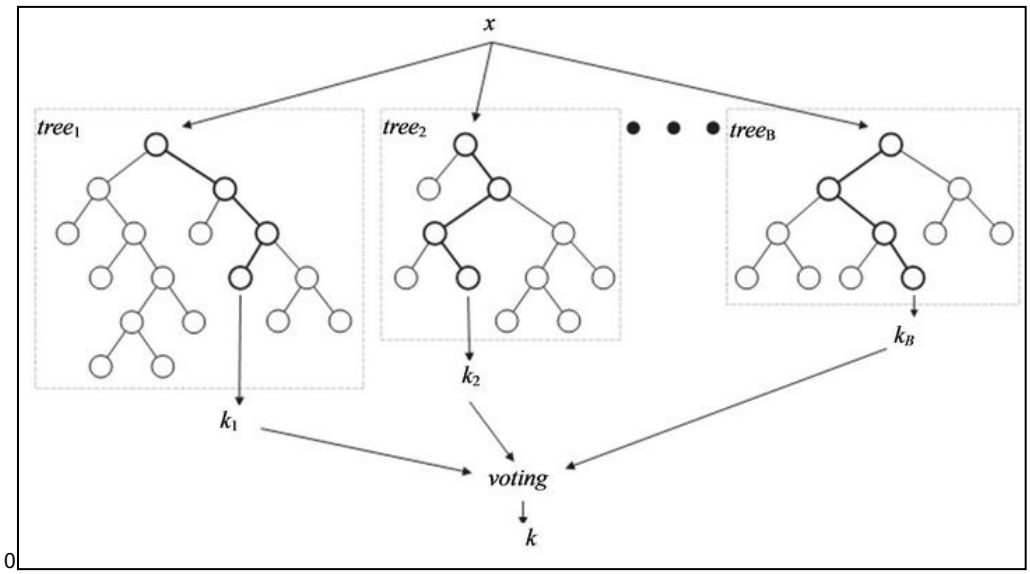
Figura 4 - Estrutura de uma árvore de decisão



Fonte: Adaptado de Chollet (2018).

No *random forest*, as variáveis são escolhidas aleatoriamente e cada nó é dividido usando o melhor entre um subconjunto de preditores escolhidos também de forma aleatória naquele nó. A partir do método de *bagging*, um método de agrupamento, as árvores sucessivas não dependem das árvores anteriores. Cada árvore é construída de forma independente usando uma amostra aleatória do conjunto de dados. Ao final, a maioria dos votos é encaminhada para a predição (BREIMAN, 1996). Um exemplo do funcionamento do algoritmo está demonstrado na Figura 5.

Figura 5 - Exemplo do funcionamento do algoritmo Random Forest



Fonte: Zhang, Cao e Romagnoli (2018).

É um algoritmo que aplica técnicas para fazer crescer muitas árvores de classificação, com a maior extensão possível, sem poda. É indicado em casos de dados com ruídos ou faltantes, quando há uma grande quantidade de variáveis (BIAU; SCORNET, 2016), com integração de bases de dados e existe a necessidade de ponderá-los e, também, em problemas de alta dimensão e com características relacionadas, como aqueles que ocorrem na área da saúde (YANG et al., 2009).

O *random forest* segue regras específicas para crescimento e combinação das árvores de decisão e é considerado estável na presença de outliers (SARICA; CERASA; QUATTRONE, 2017). Demanda pouca parametrização, sendo os principais parâmetros para sua aplicação o número de árvores nas florestas e o número de atributos, que determina o número de variáveis a serem consideradas em cada ponto de divisão das árvores (MARSLAND, 2015).

É importante ressaltar que o algoritmo se destaca não somente pela alta acurácia de suas predições, mas também por trazer informações sobre a importância das variáveis. As medidas de importância das variáveis mais utilizadas são a importância da permutação e a importância da impureza, também conhecida como importância de Gini, ou, coeficiente de Gini (BREIMAN, 2001). Ambos os métodos têm seus benefícios e limitações. A importância de Gini tende a vieses quando o conjunto de dados contém variáveis com inúmeros pontos de divisões possíveis, como no caso de variáveis categóricas ou contínuas. Por outro lado, a importância da permutação demanda bastante recursos computacionais (NEMBRINI; KÖNIG; WRIGHT, 2018).

Na importância da permutação, uma variável é identificada como importante se tiver um efeito positivo da performance preditiva, estimado pelo erro de predição OOB, *out of bag*, que compreende o erro médio de previsão em cada amostra de treinamento. A partir das predições OOB de todas as árvores na floresta, é calculado o erro quadrado médio, de acordo com a equação 1, em que Z_i é o valor medido da variável e Zi^{OOB} é a média das predições OOB (BREIMAN, 2001).

$$MSE_{OOB} = \pi r^2 = n - ^1 \sum_{i=1}^{n} [(Z_i - Zi^{OOB}])^2 \quad (1)$$

Já na importância de Gini, o coeficiente é a medida de quanto cada variável contribui para a homogeneidade de nós. A cada utilização de uma variável para divisão de um nó, os coeficientes de novos nós são calculado e comparados com o nó original (BREIMAN, 2001). O coeficiente é definido pela equação 2, sendo p=$p_i \dots p_c$ é a proporção das amostras da classe p_c para o nó m.

$$I_{G(m)} = 1 - \sum_{i=1}^{C} p_i(m^2) \quad (2)$$

A literatura apresenta trabalhos que demonstram o potencial do algoritmo *random forest* para predição e para avaliação da importância das variáveis. Luo e colaboradores (2020) destacam o uso do algoritmo *random forest* na predicação de pacientes com alto custo com doença pulmonar obstrutiva crônica. Com uma área sob a curva ROC de 0.792, os autores exploraram diferentes conjuntos de variáveis para avaliar o desempenho do modelo e, assim, por meio do coeficiente de gini, identificar aquelas que mais contribuem para a predição. Foram utilizadas 54 variáveis agrupadas em quatro categorias: demográficas, informações de doenças, informações de hospitalizações e informações sobre medicamentos.

O random forest também foi empregado no modelo preditivo proposto por Mohnen e colaboradores (2020). Com o enriquecimento do conjunto de dados administrativos utilizando informações sobre habitação, os autores avaliaram se as condições de moradia tinham impacto no custo médico de pacientes holandeses. Com uma amostra de 207.614 pacientes o modelo foi construído com a utilização de dados administrativos, dados de medicamentos, sociodemográficos e dados de um

questionário sobre condições de habitabilidade, com os seguintes subitens: habitação, espaço público, instalações públicas, composição dos habitantes (etnia, idade, tamanho da família, estabilidade) e segurança pública.

Foram construídos modelos utilizando a população geral e a população crônica para avaliação da contribuição das variáveis, que foram incluídas e excluídas para avaliação do desempenho preditivo. Os autores revelaram que a qualidade do bairro era mais importante do que sua localização, sendo tão importante quanto a idade, quando incluídas as variáveis sociodemográficas e de vizinhança. Entretanto, quando incluídas as variáveis de custos anteriores e medicamentos, a importância da qualidade do bairro diminuiu, enquanto a da idade aumentou, demonstrando que as variáveis sobre habitabilidade contribuíram apenas quando adicionadas às variáveis sociodemográficas. Neste trabalho, os autores destacaram a relevância do algoritmo *random forest* para avaliação da contribuição das variáveis (MOHNEN et al., 2020).

No trabalho desenvolvido na Coréia do Sul por Kim e Park (2019), foram incorporados aos dados administrativos três categorias de dados de check-up médico: exames laboratoriais, histórico médico e dados de comportamento autorreferidos para construir modelos de predição de pacientes com alto custo utilizando o *random forest*, regressão logística e redes neurais. A amostra compreendia pacientes entre 49 e 89 anos, considerando a elegibilidade ao check-up médico. A métrica utilizada para avaliação do modelo foi a área sob a curva ROC que, a cada incorporação de categorias de dados, teve seus resultados melhorados e mais próximos de um.

A escolha do algoritmo *random forest* se deu pela sua performance, fácil parametrização e pela possibilidade de avaliar as variáveis importantes para a classificação. Os autores demonstraram que a codificação da doença nas guias de procedimentos é um forte preditor. Entretanto, pela possibilidade de erros na inclusão desta informação, seja pela classificação incorreta, seja porque quem insere a informação é alguém responsável somente pela cobrança daquele procedimento, foi possível evidenciar a importância de outros dados na construção de modelos preditivos (KIM; PARK, 2019).

O algoritmo *random forest* se destaca na literatura devido à sua flexibilidade e precisão de previsão. Pode lidar com inúmeras variáveis e de vários tipos, com a colinearidade e com a assimetria dos dados, conforme evidenciado em um estudo comparativo que utilizou 21 algoritmos para predizer o custo de pacientes com distúrbios mentais (SHRESTHA et al., 2018).

4 ENCAMINHAMENTOS METODOLÓGICOS

Nesta seção são apresentados os encaminhamentos metodológicos que embasaram esta pesquisa. Em primeiro lugar, serão apresentadas a natureza, a população e o cenário, seguidos pelas etapas da pesquisa.

4.1 NATUREZA DA PESQUISA

Com o objetivo desenvolver um modelo de identificação de pacientes de alto custo por meio de estratégias de inteligência artificial, trata-se de uma pesquisa quantitativa, retrospectiva e de caráter descritivo. Quantitativa por realizar análises de dados sem interferência do pesquisador, retrospectiva, pois busca explorar fatos que já ocorreram e, descritiva, uma vez que busca explorar a relação entre os atributos objetos do estudo (DYNIEWICZ, 2014).

4.2 POPULAÇÃO DA PESQUISA

A população da pesquisa é composta por 586 pacientes titulares de um plano de saúde coletivo empresarial que responderam um questionário de autoavaliação de saúde.

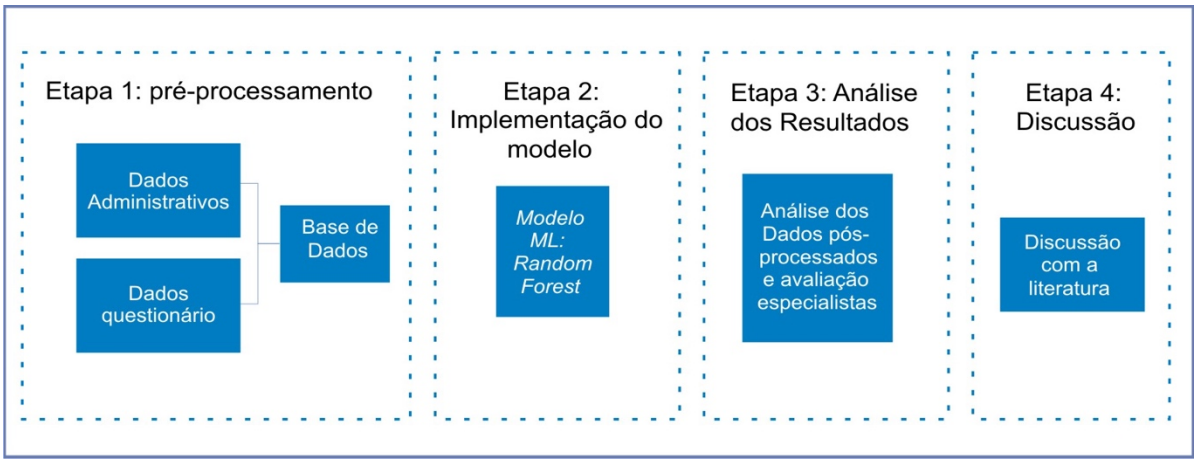
4.3 CENÁRIO DA PESQUISA

O cenário desta pesquisa é uma empresa contratante de um plano de saúde coletivo empresarial.

4.4 ETAPAS DA PESQUISA

Esta pesquisa foi realizada em quatro etapas, descritas no decorrer desta seção conforme a Figura 6.

Figura 6 - Etapas da pesquisa



Fonte: A autora (2022).

A etapa 1 compreendeu a análise de ambas as bases de dados, com o objetivo de preparar uma única base, por meio da seleção de atributos de interesse para servirem de entrada para o algoritmo (ALZUBI; NAYYAR; KUMAR, 2018). As bases utilizadas para a construção deste modelo preditivo foram as bases de dados administrativos da operadora de plano de saúde contratada pela empresa, do período de janeiro de 2019 a março de 2021 e os dados de questionário de saúde, aplicado em outubro de 2020, por meio de um link enviado aos pacientes por e-mail, e que ficou disponível para resposta por 15 dias.

Foram excluídos os registros incompletos no questionário e daqueles pacientes com histórico de neoplasias, por não ser o objetivo desta pesquisa identificar pacientes com câncer e com alto custo, uma vez que há diferenças no padrão de custo destes pacientes, com picos de custos ocorrendo particularmente após o diagnóstico e nos últimos anos de vida (DE OLIVEIRA et al., 2017).

Os dados administrativos de demandas por procedimentos médicos compreendiam: identificação (ID), data de adesão ao plano de saúde, data de nascimento, sexo, estado civil, idade, data do atendimento, competência (mês de pagamento da conta médica), internação, data de internação, data de alta, código da classificação internacional de doenças, código da tabela unificada da saúde suplementar, que padroniza nomenclaturas e códigos de procedimentos médicos (ANS, 2021b), evento (descritivo do procedimento realizado), tipo de serviço (qual categoria o procedimento é classificado: consulta, consulta em pronto-socorro, exames, terapias e internações) e valor pago pelo procedimento em reais.

Com estes dados, foram criadas variáveis de quantidade de procedimentos realizados, conforme trabalhos propostos Bertsimas et al. (2008), Duncan, Loginov e Ludkovski (2016) e colaboradores, distribuídas nas seguintes categorias, já pré-definidas na base de dados pela própria operadora de plano de saúde: consultas, composta por todas as consultas descritas como eletivas, com clínico ou especialista, consultas em pronto-socorro, que compreendem todas as consultas realizadas em regime de urgência e/ou emergência, exames, internações, realizadas em regime de hospital ou dia ou não, e terapias, em que se enquadram procedimentos como fisioterapia, psicoterapia ou terapia ocupacional. Um exemplo de como foi realizado este quantitativo, encontra-se no Quadro 3.

Quadro 3 - Exemplo da apuração da quantidade de procedimentos

Codigo_evento	Evento	Categoria	Quantidade
1.01.01039	Consulta em pronto socorro	Consulta em pronto socorro	1
4.03.02040	Glicose - pesquisa e/ou dosagem	Exames	1
4.03.02733	Hemoglobina glicada (fração A1C) - pesquisa e/ou dosagem	Exames	1
2.01.03492	Patologia osteomioarticular em dois ou mais membros	Terapia	1
5.00.00470	Sessão de psicoterapia individual por psicólogo	Terapia	1
1.01.02019	Visita hospitalar paciente internado	Internação	1

Fonte: A autora (2022).

O custo total de todas as demandas por procedimentos de saúde destes pacientes foi apurado para poder realizar a classificação daqueles com alto custo e baixo custo. Foram classificados como alto custo aqueles pacientes que compreendiam 50% do custo total.

Devido à pandemia da covid-19 e, por consequência, a diminuição na demanda por procedimentos médicos eletivos, foram criadas variáveis a fim de avaliar se houve impacto da pandemia nas demandas por procedimentos destes pacientes considerados alto custo, com variáveis dicotômicas relacionadas aos custos terem incorridos antes de 2020, após 2020 ou em ambos os períodos.

As variáveis selecionadas a partir da base de dados administrativos encontram-se no Quadro 4.

Quadro 4 - Variáveis da base de dados administrativos

quantidade de consultas eletivas	total de consultas realizadas no período
quantidade de consultas em pronto-socorro	total de consultas em pronto-socorro realizadas no período
quantidade de exames	total de exames realizados no período
quantidade de internações	total de internações realizadas no período
quantidade de terapias	total de terapias realizadas no período
classificação de custo	custo total de 2019-2021 para classificação em alto e baixo custo
custos até 2020	se o paciente teve custos até 2020
custos após 2020	se o paciente teve custo após 2020
custos ambos os períodos	se o paciente teve custo em ambos os períodos

Fonte: A autora (2022).

Os dados autorreferidos que caracterizam a percepção dos pacientes sobre sua saúde foram obtidos por meio de questionário que foi respondido por meio de um link. O questionário tinha como objetivo realizar um mapeamento de saúde dos pacientes para o planejamento e execução de ações de promoção a saúde na empresa. O questionário encontra-se no Anexo B. Esta base de dados continha a mesma identificação (ID) da base de dados administrativo, o que facilitou a criação da base de dados final com todas as variáveis de interesse.

Após avaliação das questões e respostas oriundas do questionário, foram elencadas aquelas com potencial relação com o desfecho alto custo (ALZUBI; NAYYAR; KUMAR, 2018; WAMMES et al., 2018; SMITH et al., 2019) e aquelas relacionadas a prontidão para mudança, para buscar entender o desafio dos pacientes perante suas condições, fator importante para inclusão, gestão e êxito de pacientes em programas de saúde (ZULMAN et al., 2015; SMITH et al., 2019). Assim, foram criadas as variáveis apresentadas no Quadro 5.

Quadro 5 - Variáveis do questionário de avaliação de saúde

continua

Nome	Resultado	Observações
Idade	Idade em anos	Idade no momento da resposta ao questionário
Sexo	Masculino e feminino	
IMC	Categorização segundo OMS	Peso em kg sobre a altura ao quadrado
Qualidade de Vida	Muito satisfeito, satisfeito, neutro, insatisfeito, muito insatisfeito	Avaliar a satisfação com a qualidade de vida
Consultas	Sim ou não	Avaliar a realização de consultas regulares/anuais
Vacinas	Sim, não ou não sei	Avaliar a realização de vacinas
Medicamento	Sim ou não	Avaliar o uso diário de medicamento
Hipertensão Arterial	Sim ou não	Identificar pacientes hipertensos
Hipertensão Arterial Medicamentos	Sim ou não	Identificar pacientes hipertensos em terapia medicamentosa
Diabetes Melitus	Sim ou não	Identificar pacientes diabéticos
Diabetes Melitus Medicamentos	• Sim, comprimidos; • sim, insulina; • não.	Identificar pacientes diabéticos em terapia medicamentosa
Hipercolesteremia	Sim ou não	Identificar pacientes com colesterol alto
Hipercolesteremia Medicamentos	Sim ou não	Identificar pacientes com colesterol alto em terapia medicamentos
Estado de saúde mental	Ótimo, bom, regular e ruim	Avaliar a autopercepção da saúde mental
Acompanhamento psicológico	Sim ou não	Identificar pacientes em acompanhamento psicológico
Acompanhamento psiquiátrico	Sim ou não	Identificar pacientes em acompanhamento psiquiátrico
Uso de medicamento controlado	Sim ou não	Identificar pacientes em uso de medicação controlada
Horas de Sono	• 4 horas ou menos • 4 a 6 horas • 7 horas ou mais	Avaliar quantidade de sono diário
Tabagismo	Sim ou não	Identificar pacientes tabagistas
Elitismo	Sim ou não	Identificar pacientes em uso frequente de álcool
Atividade física	Sim ou não	Identificar paciente ativos fisicamente
Condição de saúde e atividade física	Sim ou não	Identificar aqueles pacientes cuja condição de saúde impedem a realização de exercício

Quadro 5 - Variáveis do questionário de avaliação de saúde

conclusão

Nome	Resultado	Observações
Motivação par mudar hábitos (atividade física)	• Não se aplica • Não penso nisso • Penso um pouco em mudar • Penso muito em mudar • Estou tomando atitudes para mudar	Avaliar a prontidão para mudança de hábitos relacionados a atividade física
Quantidade de refeições	De 3 a 6 refeições diárias	Identificar a quantidade de refeições feitas diariamente
Motivação para mudar hábitos (refeição)	• Não se aplica • Não penso nisso • Penso um pouco em mudar • Penso muito em mudar • Estou tomando atitudes para mudar	Avaliar a prontidão para mudança de hábitos relacionados a alimentação
Dor forte	Sim ou não	Identificar pacientes com dor crônica
Histórico pessoal de doenças (epilepsia, obesidade, problemas de visão, alergias de pele, doenças de tireoide, doenças do estômago, rinite, sinusite, asma, hepatite B, hepatite C ou HIV, dor de cabeça, insônia, ansiedade, infarto, insuficiência cardíaca crônica, depressão, anemia, doença renal, artrose, doença pulmonar obstrutiva crônica, transtorno bipolar, transplantes de órgãos)	Sim ou não	Identificar pacientes com histórico de doenças
Histórico familiar de doenças (diabetes, colesterol e/ou triglicerídeos alto, acidente vascular cerebral)	Sim e não	Identificar pacientes com histórico familiar de doenças

Fonte: A autora (2022).

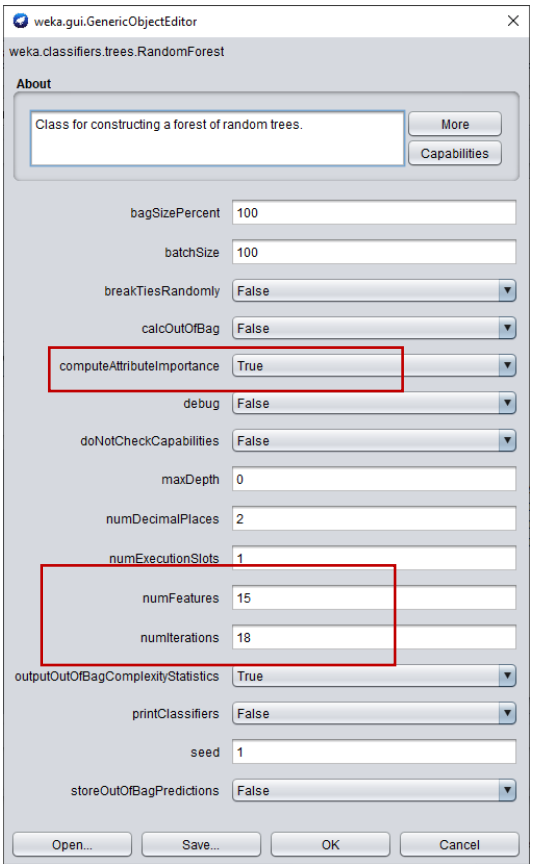
A idade considerada para a inclusão no modelo preditivo foi a idade no momento da resposta ao questionário. Com os dados de peso e altura foram calculado os índices de massa corpórea e, em seguida, classificados conforme recomendação da Organização Mundial da Saúde para o diagnóstico do estado nutricional (OMS, 2000): magreza (IMC < 18,5), adequado (IMC entre 18,5 e 24,9), sobrepeso (IMC entre 25,0 e 29,9), obesidade grau I (IMC entre 30,0 e 34,9), obesidade grau II (IMC entre 35,0 e 39,0) e obesidade grau III (IMC > 40,0).

As variáveis sobre histórico pessoal de doenças foram criadas a partir de uma pergunta fechada, em que o respondente deveria assinalar quais doenças apresenta

ou apresentou. Com as respostas, foram criadas variáveis para cada doença passível de resposta no questionário.

Após a preparação do conjunto de dados, a etapa 2 foi seguida para a aplicação do algoritmo Random Forest, no WEKA, *Waikato Environment for Knowledge Analysis* (HALL et al., 2009). Foram definidos os parâmetros relacionados ao número de árvores (*numInterations*) e o número de atributos (*numFeatures)*. O valor definidos para o número de árvores foi 18 e para o número de variáveis foi 15. Estes valores foram encontrados após diversas simulações, seguindo o descrito por Marsland (2015), bastando aumentar o número de árvores até que o valor do erro pare de aumentar. Para auxiliar na interpretação dos resultados, foram selecionadas para que as importâncias dos atributos também saíssem como resultado, no campo *computeAttributeIMportance.* A tela de configuração do algoritmo encontra-se na Figura 7.

Figura 7 - Configuração dos parâmetros do algortimo *Random Forest*



Fonte: A autora, adaptado do Weka® (2022).

Devido ao tamanho da amostra, foi utilizado o método de validação cruzada em 10 partições para a etapa de treinamento e teste (DOUPE; FAGHMOUS; BASU, 2019).

Na etapa 3, composta pela análise dos dados pós-processados, foram avaliados os resultados considerando as métricas para algoritmos de classificação proposta por Sokolova e Lapalme (2009): acurácia, sensibilidade, especificidade e área sob a curva ROC. Também foi realizada uma análise da importância das variáveis por meio da importância de gini (BREIMAN, 2001; NEMBRINI; KÖNIG; WRIGHT, 2018).

Após a estruturação dos resultados, em que foram trazidas as métricas e elencadas as variáveis por grau de importância, foram encaminhados para especialistas um questionário de avaliação, com o objetivo de medir o seu grau de concordância em relação à importância das variáveis para a construção do modelo e a contribuição de um modelo com resultados auditáveis para a gestão de saúde.

Como critérios de inclusão do especialista participante da pesquisa, o profissional deveria possuir no mínimo dois anos de experiência em gestão de saúde. Para garantir um olhar multidisciplinar a seleção envolveu a busca por médicos, enfermeiros e gestores. Não foram selecionados profissionais que atuam na instituição onde os dados foram coletados. A recomendação de especialistas sofre variações na literatura e, para esta pesquisa, foi seguida a recomendação proposta por Lynn (1986), que define um número mínimo de cinco e máximo de dez.

A busca dos especialistas foi realizada por meio de amostragem não probabilística, denominada bola de neve. Esse método é aplicado quando o acesso a especialistas com características específicas é considerado restrito. No método bola de neve, os especialistas com características desejáveis para o estudo recrutam futuros especialistas entre sua rede de contatos (NADERIFAR; GOLI; GHALJAIE, 2017; SEDGWICK, 2013). Mediante contato com um gestor de saúde, foi feita uma explicação sobre os objetivos da pesquisa e solicitada a indicação dos demais profissionais para participar e seus respectivos endereços eletrônicos.

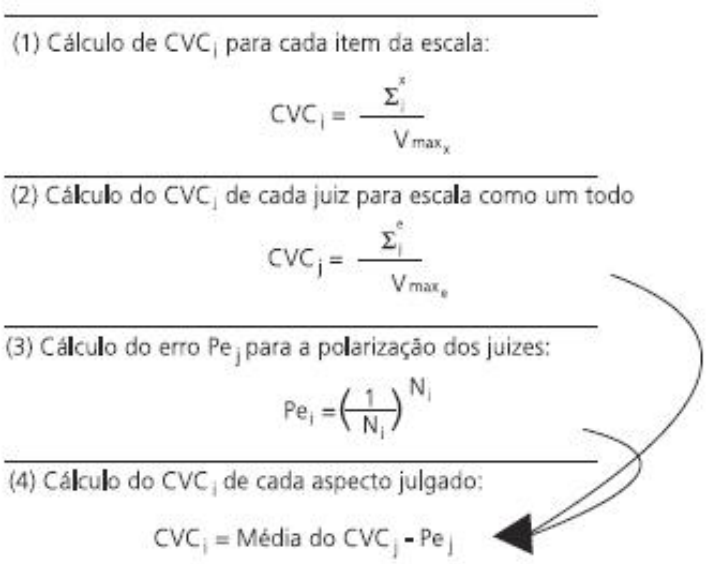
Considerando o cenário pandêmico, a comunicação e envio da avaliação foram feitos eletronicamente, com informações sobre o objetivo da pesquisa, o papel do especialista nesta etapa, o tempo estimado para o preenchimento do formulário, o *link* de acesso, o prazo para preenchimento do formulário e o Termo de Consentimento Livre e Esclarecido (TCLE), cujo documento encontra-se no Apêndice A. O questionário ficou habilitado para respostas por 10 dias, até que foi encerrado pela pesquisadora.

O questionário foi construído em ferramenta *Google Forms*, para facilitar o envio aos especialistas por meio eletrônico e foi composto por 11 seções, disponível no Apêndice B. A Seção 1 trazia o contexto da pesquisa, agradecimentos aos respondentes e o tempo estimado de resposta. Na seção 2 foi disponibilizado o TCLE para que, na seção 3, o especialista concordasse ou não em participar da pesquisa e, na seção 4, seu e-mail era solicitado, para envio do TCLE, em caso de concordância ou não. Na condição de aceite, o questionário seguia para as seções 5, 6 e 7, com questionamentos sobre a atuação em gestão em saúde, o tempo de atuação (se inferior a 2 anos, o questionário era encerrado) e formação acadêmica.

Na seção 8 foi disponibilizada um resumo dos objetivos da pesquisa, método e resultados, para que o especialista pudesse entender o contexto e principais descobertas da pesquisa. Na seção 9, que envolveu a avaliação do grau de importância das variáveis para a construção do modelo e sua correlação com o desfecho alto custo, foram disponibilizadas as variáveis de acordo com o grau de importância de gini e solicitado que os especialistas assinalassem cada afirmativa de acordo com a sua concordância para a alternativa exposta. O estabelecimento da concordância de cada questão utilizou a proposta da escala de Likert (LIKERT, 1932), com variação de 1 a 5, sendo: concordo totalmente, concordo, indiferente, discordo e discordo totalmente. Em caso discordância, estava disponível um campo para justificativa.

Os resultados foram exportados para planilha e utilizados para calcular o Coeficiente de Validade de Conteúdo (CVC), um método de análise muito utilizado na área da saúde que possui como principal finalidade medir a proporção ou a porcentagem de especialistas que estão em concordância sobre os aspectos relacionados ao estudo analisado (VIEIRA et al., 2020). O cálculo do CVC foi realizado com base no proposto por Hernández Nieto (2002), transcrito no trabalho de Filgueiras e colaboradores (2015), e sua equação encontra-se na Figura 8.

Figura 8 - Cálculo do coeficiente de validade de conteúdo



Fonte: Filgueiras et al. (2015).

No que se refere a taxa de concordância aceitável, autores destacam que devem ser considerados o número de participantes, e recomenda um CVC igual ou superior a 0,78 para seis ou mais deles (ALEXANDRE; COLUCI, 2011; VIEIRA et al., 2020) o qual foi adotado para esta pesquisa.

Por fim, a etapa 4 compreendeu a discussão dos resultados encontrados com a literatura.

4.5 ASPECTOS ÉTICOS

Os dados utilizados neste estudo já haviam sido coletados e analisados para outros fins que não o desta pesquisa sendo, portanto, oriundos de bases de dados secundários. Para a utilização dos dados e para garantir a participação dos especialistas, o projeto de pesquisa foi submetido ao Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Paraná sob o e aprovado sob o parecer número 4.756.053 (Anexo A).

5 RESULTADOS

Nesta seção serão apresentados os resultados obtidos, trazendo as decorrências do modelo preditivo e da avaliação dos especialistas.

Como resultado do pré-processamento foi obtido um único conjunto de dados com dados administrativos e autorreferidos dos pacientes, que totalizaram 63 variáveis, descritas previamente nos encaminhamentos metodológicos. Por se tratar de bases de dados anonimizadas, os pacientes estavam identificados por um identificador (ID), que permitiu criar a vinculação entre as bases de dados em uma única planilha.

A população deste estudo foi composta por 586 pacientes. O questionário de saúde foi aplicado por meio de um link e alguns pacientes iniciaram e não concluíram, registrando apenas os dados demográficos, e outros preencheram e salvaram mais de uma vez. Por este motivo, foram feitas as exclusões daqueles registros sem respostas e daqueles duplicados, reduzindo o número de pacientes daquela amostra inicial recebida (n: 30). Também foram excluídos aqueles pacientes com histórico de câncer (n: 3), chegando a uma amostra de 553 pacientes.

Em relação aos dados demográficos, 74,5% são do sexo feminino (n: 412) e 25,5% do sexo masculino (n: 141), com idade média de 39 anos. O paciente mais novo tem 17 anos e o mais velho 69 anos. Por haver obrigatoriedade no preenchimento de todos os campos do questionário, não foram identificados valores faltantes.

Ao realizar a categorização do índice de massa corpórea, ficou evidenciado o estado nutricional desta população, com mais de 50% com sobrepeso ou algum nível de obesidade, conforme Tabela 1.

Tabela 1 – Frequência relativa de pacientes segundo estado nutricional

IMC categorizado	Percentual de pacientes
adequado	31
magreza	1
sobrepeso	40
obesidade_I	20
obesidade_II	5
obesidade_III	3

Fonte: A autora (2022).

Dos 46 pacientes classificados como alto custo, apenas 7 estão com o IMC considerado adequado, 21 estão com sobrepeso, 13 têm obesidade grau I, 4 têm obesidade grau II e 1 tem obesidade grau III.

Com relação às doenças crônicas não transmissíveis, a prevalência foi: 18% de hipertensos e 5,4% de diabéticos. Com relação às outras doenças autorreferidas, destacaram-se os transtornos mentais, com uma prevalência de 16,3% para ansiedade e 3,6% para depressão.

O custo total do atendimento médico destes pacientes foi de R$ 5.727.735,32 ao longo do período. Para a classificação dos pacientes com alto custo, foram considerados aqueles que correspondem a 50% do custo total. Estes pacientes representam 8,3% do total (n: 46).

A estatística descritiva dos valores gastos pelos dos pacientes classificados como alto custo encontra-se na Tabela 2. O paciente com o maior custo teve um custo assistencial no período de R$ 257.599,93 e a média de custo destes pacientes é de R$ 62.257,34.

Tabela 2 - Estatística descritiva dos valores gastos por pacientes de alto custo

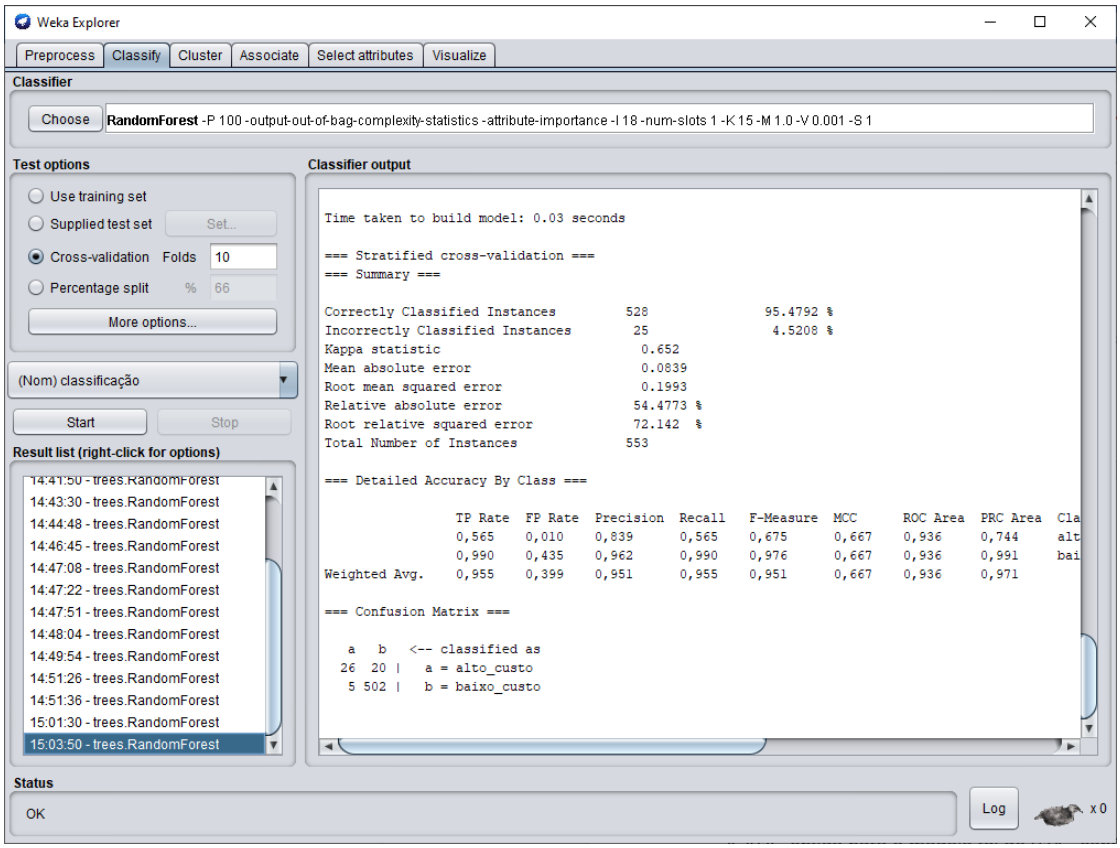
Estatística Descritiva	Valores (R$)
Média	62.257,34
Erro Padrão	5.203,65
Mediana	52.617,81
Desvio padrão	35.292,86
Mínimo	48.166,69
Máximo	257.599,93
Soma	2.863.837,66

Fonte: A autora (2022).

5.1 MODELO PREDITIVO PARA IDENTIFICAÇÃO DE PACIENTES DE ALTO CUSTO

Após a inclusão do conjunto de dados no software WEKA e selecionado o algoritmo *random forest*, foram obtidos os seguintes resultados, conforme tela de resultados apresentada na Figura 9. Os resultados de acurácia, sensibilidade e especificidade encontram-se na sequência, na Tabela 3.

Figura 9 - Tela da aplicação do algoritmo *Random Forest* no Weka



Fonte: A autora, adaptado do Weka® (2022).

Tabela 3 - Resultados obtidos de acurácia, sensibilidade e especificidade utilizando o algoritmo *random forest*

Modelo	Acurácia	Sensibilidade	Especificidade
Random Forest	95,5%	95,1%	95,5%

Fonte: A autora (2022).

A matriz confusão obtida para o modelo encontra-se na Tabela 4:

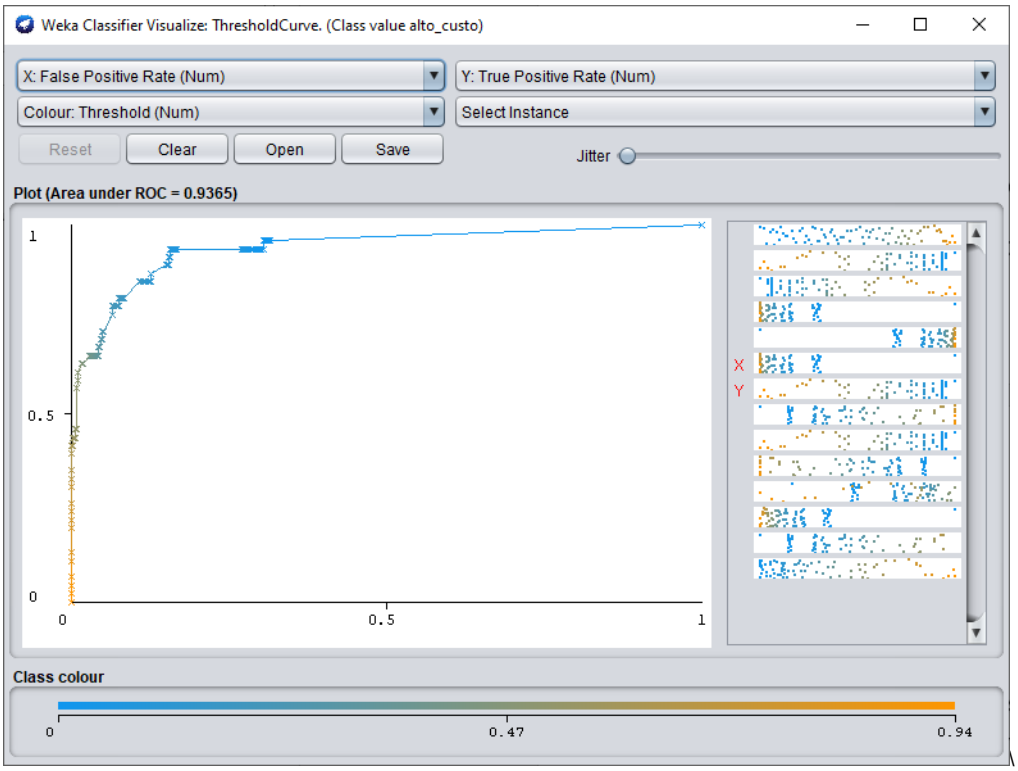
Tabela 4 - Matriz confusão para o algoritmo *random forest*

	VP	FP
VN	26	20
FN	5	502

Fonte: A autora (2022).

A AUC obtida para o modelo foi de 0,93, conforme demonstrado na Figura 10.

Figura 10 - AUC – Área sob a curva ROC



Fonte: A autora, adaptado do Weka® (2022).

No Quadro 6 encontram-se as listadas das variáveis avaliadas pelo grau de importância pelo método de importância de Gini, cujo resultado é a medida de quanto cada variável contribui para a homogeneidade dos nós.

Quadro 6 – Grau de Importância das Variáveis

continua

Importância de Gini	Variáveis
1.00	pessoal_transtorno_bipolar
0.8	medicamento
0.73	refeicoes_realizadas_diariamente
0.66	sente_dor_forte
0.64	vacinas_atualizadas
0.61	sexo
0.59	uso_bebida_alcoolica
0.59	pessoal_ansiedade
0.56	pessoal_depressao
0.55	condicao_fisica_impede_realizar_exercicio
0.55	fez_acompanhamento_com_psicologo
0.52	estado_de_saude_mental
0.52	tabagista

Quadro 6 - Importância de Gini

conclusão

Importância de Gini	Variáveis
0.52	idade
0.48	familiar_diabetes
0.44	IMC_cat
0.43	familiar_colesterol_e_ou_triglicerideos_alto
0.43	motivacao_mudar_habito_atividade_fisica
0.39	quantas_horas_costuma_dormir_por_noite
0.38	procedimentos_terapia
0.38	faz_uso_de_medicamento_controlado
0.37	qualidade_de_vida
0.36	hipercolesterolemia
0.35	pessoal_obesidade
0.35	procedimentos_exames
0.34	hipertensao
0.32	consultas_regulares
0.31	pessoal_intestino
0.31	realiza_atividade_fisica
0.30	procedimentos_internacao
0.28	procedimentos_c
0.28	pessoal_alergia_pele
0.28	procedimentos_cps
0.27	motivacao_mudar_habito_alimentacao
0.16	diabetes
0.15	pessoal_insonia
0.14	custos_apos_2020
0.14	pessoal_rim
0.13	custos_ate_2020
0.13	pessoal_tireoide

Fonte: A autora (2022).

5.2 AVALIAÇÃO DOS ESPECIALISTAS

As questões presentes no questionário foram respondidas por especialistas em gestão de saúde. Ao todo, participaram 6 especialistas, sendo 5 médicos e 1 enfermeiro. Todos atuavam com gestão em saúde há mais de dois anos, sendo 3 dos especialistas com 10 anos ou mais de atuação, 2 com 3 a 4 anos de atuação e 1 com 5 a 10 anos de atuação.

As respostas dos especialistas foram analisadas com o objetivo de avaliar se o grau de importância das variáveis incorridas pelo algoritmo estão relacionadas com o desfecho alto custo. Também foi questionado sobre qual seria a melhor opção para apoio a decisão em gestão de saúde, um algoritmo interpretável com performance inferior, ou um algoritmo com performance superior, porém não interpretável.

Na Tabela 5 encontram-se as variáveis avaliadas, o número de respostas obtidas e o respectivo resultado do CVC. Das dez variáveis avaliadas, duas apresentaram CVC acima de 0,90 (dores fortes e uso de bebida alcóolica), seis ficaram entre 0,78 e 0,89 (uso contínuo de medicamentos, quantidade de refeições realizadas diariamente, histórico de vacinas, histórico pessoal de ansiedade e histórico pessoal de depressão), e duas resultaram em CVC inferior a 0,78 (transtorno bipolar e acompanhamento com psicólogo). Em média, o questionário apresentou um CVC de 0,81.

Tabela 5 - Variáveis avaliadas, respostas obtidas e CVC

Variáveis	Número de respostas	CVC
Transtorno bipolar	6	0,73
Uso contínuo de medicamentos	6	0,80
Quantidade de refeições realizadas diariamente	6	0,80
Dores fortes	6	0,93
Histórico de vacinas	6	0,80
Uso de bebida alcóolica	6	0,90
Histórico pessoal de ansiedade	6	0,80
Histórico pessoal de depressão	6	0,87
Condição física impedir de realizar exercícios físicos	6	0,87
Fazer acompanhamento com psicólogo	6	0,60
	Média Final do CVC	**0,81**

Fonte: A autora (2022).

O grau de concordância com as variáveis fazer acompanhamento com psicólogo e ter transtorno bipolar foram analisadas separadamente, por apresentarem um CVC abaixo do grau de aceitação de 0,78, como preconizado pela literatura (ALEXANDRE; COLUCI, 2011; VIEIRA et al., 2020). Em relação à variável transtorno bipolar, 5 especialistas concordaram com a associação desta variável com o desfecho alto custo e um especialista discordou. Uma vez que foram disponibilizados campos para justificativa quanto à discordância, a seguir encontra-se o comentário do especialista um, relacionado a sua discordância:

> *Penso que pelo diagnóstico não seria um custo com plano de saúde. Me faz pensar que a condição mental do indivíduo adoecido que procura muitos serviços de saúde, tem um comportamento particular, muitos procedimentos e exames poderiam o levar a ter um diagnóstico de TAB pelo excesso de médicos. Por ser uma população pequena, não associaria* [diz o especialista 1, que discordou].

Em relação à variável fazer acompanhamento psicológico, dois especialistas concordaram, dois consideraram indiferente e dois discordaram. Em relação às justificativas, destacam-se:

> *Penso ser parte de um tratamento mais complexo como um todo. E para lidar com tudo isso, há o encaminhamento ao psicólogo para lidar com a situação* [diz o especialista 1, que discordou].

> *Entendo que fazer acompanhamento psicológico regular ajudaria na redução do alto custo* [diz o especialista 5, que discordou].

Apesar do CVC de 0,80 para a variável quantidade de refeições realizadas diariamente, obtido por meio de 3 resultados de concordância, 2 de concordância total e 1 de discordância, dois especialistas comentaram:

> *Isoladamente, não veria muita associação sem mais dados da saúde do indivíduo (diagnósticos)* [diz o especialista 1, que discordou].

> *Depende se a refeição for fracionada, composição e outros dados para que a associação com desfecho de alto custo possa ser analisada melhor* [diz o especialista 6, que concordou].

Outra associação que obteve comentário de um especialista foi a relacionada ao uso contínuo de medicamentos. Com CVC de 0,80, o único especialista que discordou, destacou:

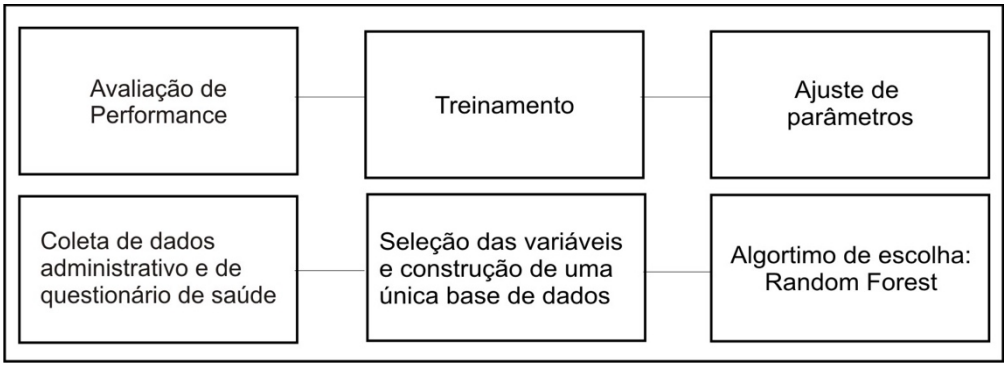
> *A quantidade, indicando uma polifarmácia, sim. Mas não o uso de medicação contínua isoladamente* [diz o especialista 1].

A aplicação do questionário aos especialistas permitiu a avaliação das variáveis mais importantes para a construção do modelo preditivo, de acordo com a importância de gini. Ainda que não tenham sido avaliadas as associações entre elas, os resultados permitiram o enriquecimento das análises dos resultados e, ainda, podem aperfeiçoar uma proposta de questionário futura.

Quando questionados sobre o que é mais desejável como elemento de apoio à decisão em gestão de saúde: uma melhor performance do algoritmo, indicando com maior precisão quem serão os pacientes com risco de se tornarem alto custo, sem explicação de como chegou ao seu resultado, ou uma performance ligeiramente inferior, em que é possível identificar quais variáveis estão mais fortemente associadas a este desfecho, todos os especialistas responderam preferir uma performance ligeiramente inferior, com a identificação das variáveis, optando, portando, por aqueles algoritmos interpretáveis para a gestão de saúde.

Diante dos resultados obtidos, cumpriu-se, então, os objetivos desta pesquisa, de identificar e propor um modelo de identificação de pacientes de alto custo com dados administrativos e de questionário de saúde, demonstrado na Figura 11, além de avaliar a importância das variáveis para a construção do modelo.

Figura 11 - Modelo de identificação de pacientes de alto custo



Fonte: A autora (2022).

Para classificar pacientes de alto custo, deve-se, incialmente, realizar a etapa de coleta dos dados e seu pré-processamento. Da base de dados administrativos serão extraídas as variáveis de quantitativo de procedimentos realizados por cada paciente em cada categoria de custo (consultas eletivas, consultas em pronto-socorro, internações, exames e terapias) e realizada a apuração do custo total. Da base de dados do questionário de saúde, são extraídas aquelas variáveis de histórico médico pessoal e familiar, saúde mental, sono, hábitos e motivação para mudança. Após esta seleção, uma única base de dados é criada com os dados de cada paciente.

Na sequência o algoritmo *random forest* é implementando com sua parametrização, definindo o número de árvores e o número de atributos. Selecionando o parâmetro para que o resultado apresente a importância das variáveis, será possível identificar aquelas mais importantes para o resultado e, a partir destas

informações, estruturar ações para eleger, monitorar e incluir pacientes em programas de saúde, para atuação de forma preventiva e curativa. A avaliação da performance do algoritmo deve ser acompanhada a fim de garantir que as classificações estejam realizadas adequadamente. Além da acurácia, a sensibilidade e a especificidade e curva ROC são métricas que devem ser avaliadas constantemente.

6 DISCUSSÃO

Neste capítulo serão discutidos com a literatura os principais resultados obtidos, elencadas as limitações desta pesquisa e as perspectivas de trabalhos futuros.

6.1 DISCUSSÃO COM TRABALHOS RELACIONADOS

O algoritmo escolhido para a construção do modelo tem desempenho relatado na literatura para problemas de classificação na área da saúde, destacando-se como um algoritmo de boa performance na presença de inúmeras variáveis e que permite a avaliação do poder preditivo destas variáveis (LUO et al., 2020; MOHNEN et al., 2020; KIM; PARK, 2019), conforme os resultados obtidos também por esta pesquisa. O modelo proposto com a utilização de dados administrativos e autorreferidos obteve resultados relevantes quando analisadas as métricas de performance de algoritmos de classificação (SOKOLOVA; LAPALME, 2009) e, portanto, permitiu um abordagem válida, confiável e implementável para predição de pacientes de alto custo, conforme sugerido na literatura (SIEKMAN; HILGER, 2018).

Os pacientes classificados como alto custo, considerados aqueles que compreendem 50% do custo total, representaram 8,31% do total de pacientes, condizente com o relatado na literatura, de que pacientes de alto custo compreendem entre 1 e 20% do total de pacientes (WAMMES et al., 2017)

Os resultados obtidos nesta pesquisa permitiram a identificação de variáveis não relacionadas a custo para predição de pacientes de alto custo, como no estudo estatístico realizado por Fleishman e Cohen (2010), em um estudo coorte, que destacou que as informações sobre as condições médicas dos pacientes melhoraram o poder preditivo, além das variáveis já estabelecidas como gênero e idade, como a quantidade de doenças crônicas, a autopercepção do estado de saúde e perda de funcionalidade.

Quando analisados os resultados da importância de Gini, foi possível constatar a importância das variáveis autorreferidas. Quanto maior a importância de Gini, mais relevante é essa variável para manter o poder preditivo do modelo com a utilização do *Random Forest* (BREIMAN, 2001). Destacaram-se neste resultado aquelas variáveis relacionadas a saúde mental, como transtorno bipolar, depressão, ansiedade e uso de álcool. A literatura traz estudos que demonstram os transtornos mentais

como característica dos pacientes de alto custo. Em estudo conduzido por Buck, Teich e Miller (2003), que analisou pacientes em acompanhamento por transtornos mentais e abuso de drogas do programa Medicaid, foi possível concluir que apesar de compreenderem apenas 11% do total de beneficiários do programa, estes respondiam por um terço daqueles pacientes com alto custo.

Em estudo transversal que utilizou dados administrativos e autorreferidos para analisar os pacientes com alto custo no Canadá (ROSELLA et al., 2014), foi possível perceber as características individuais de pacientes com alto custo, incluindo aquelas relacionadas à saúde mental: estes pacientes tinham realizado atendimento especializado com psicólogo ou psiquiatra no último ano e tinham uma pior autopercepção em relação à sua saúde mental, condizente com os resultados desta pesquisa que evidenciou o acompanhamento com psicólogo como uma variável importante para a predição de pacientes com alto custo. Os diagnósticos psiquiátricos e o uso de medicamentos psicotrópicos foram relacionados a uma maior chance de se tornar um paciente com alto custo em um estudo transversal conduzido nos Estados Unidos (DOVE; DUNCAN; ROBB, 2003).

É necessário destacar que os transtornos mentais estão presentes na população trabalhadora do Brasil, estando entre as 3 principais causas de afastamento do trabalho (MREJENS; RACHE; NUNES, 2021). Em 2020 foram concedidas mais licenças pelo Instituto Nacional do Seguro Social devido a transtornos mentais e comportamentais, num incremento de 29,1% em relação a 2019 (MREJENS; RACHE; NUNES, 2021). A despeito da concordância entre os especialistas sobre a relação destas variáveis com pacientes de alto custo, é importante destacar que o momento da coleta dos dados pode ter proporcionado uma importância maior destas variáveis.

Contudo, apesar do momento pandêmico e da diminuição da busca por procedimentos médicos neste período (SILVA et al., 2020), não foi possível relacionar associar os pacientes de alto custo a incorrência destes custos no período pré ou pós-pandemia.

A relação dos hábitos alimentares com pacientes com alto custo também está amplamente discutida na literatura e condizente com os achados nesta pesquisa. A dieta inadequada está relacionada a doenças crônicas que, por usa vez, estão relacionadas aos pacientes com alto custo (BLUMENTHAL et al., 2016). Entretanto, um maior detalhamento sobre os hábitos alimentares pode ser necessário para maior compreensão do impacto destas variáveis com o desfecho alto custo. Ainda, hábitos

alimentares podem estar associados a questões sociais (ALBERGA et al., 2018) que, nesta pesquisa, não foram consideradas, uma vez que a variável compreendia apenas o quantitativo de refeições realizadas diariamente, sem considera sua qualidade, o que foi levantado em argumentação dois dos especialistas. Da mesma forma, o uso de bebida alcóolica está associado a inúmeras doenças crônicas custosas aos sistemas de saúde como câncer, hipertensão (ALBERGA et al., 2018). Contudo, estudos demonstram que após o início do tratamento médico, o uso do álcool é diminuído (ALBERGA et al., 2018; ROSELLA et al., 2014) e, por isso, se faz necessário um melhor entendimento da relação uso de bebida alcóolica com pacientes de alto custo.

O uso contínuo de medicamentos está associado aos pacientes com alto custo. Em estudo transversal que avaliou o consumo de medicamentos e as características da população da Columbia Britânica, no Canadá, constatou que os pacientes com alto custo tinham uma maior chance de fazer uso contínuo de medicamentos. Uma pequena proporção da população é responsável pela maior parte dos gastos com medicamentos e estes padrões de gastos persistem com o tempo (WEYMANN et al., 2017). Contudo, é importante considerar o comentário feito pelo especialista 1, em relação ao paciente ser, ou não, polifarmácia.

De acordo com a Organização Mundial da Saúde, polifarmácia é o uso concomitante de quatro ou mais medicamentos, com ou sem prescrição médica, por um paciente (OMS, 2021). O uso de medicamento contínuo pode caracterizar o tratamento de doenças de baixa complexidade, que não necessariamente irão incorrer em alto custo e uma questão mais detalhada no questionário de saúde pode elucidar esta questão.

Poucos estudos trazem o impacto da dor crônica nos custos com saúde e como ela afeta a utilização dos serviços de saúde. Numa avaliação de dados administrativos para estimar o custo incremental da dor crônica em uma população de jovens adultos na província de Ontario, foi possível determinar que os custos incrementais foram mais altos para aqueles pacientes com dor intensa e naqueles com maior limitação de atividades. Os custos per capita de pacientes com dores crônicas foi 50% maior, quando comparados aos pacientes sem dores (HOGAN et al., 2016). As dores crônicas estão associadas não somente a maiores custos com atendimento médico, mas também aos custos indiretos relacionados a menor

produtividade, menos horas trabalhadas e salários perdidos por trabalhadores (GASKIN; RICHARD, 2011).

Um estudo realizado na Bahia destacou a forte relação das dores crônicas com a incapacidade laboral, com 70% dos participantes do estudo em afastamento previdenciário ou sem atividade remunerada (CASTRO et al., 2011). Considerando o cenário e população desta pesquisa, a importância da variável sentir dores fortes ganha relevância, uma vez que afeta não somente os custos com saúde, mas também a capacidade laboral dos colaboradores e a produtividade da empresa.

As variáveis provenientes dos dados autorreferidos tiveram maior importância para os resultados deste modelo. A partir destes resultados, melhorias na construção do questionário podem ser propostas, a fim de obter maiores informações referente ao consumo de medicamento, de bebida alcóolica e qualidade das refeições, por exemplo. A pergunta de múltipla escolha sobre doenças, que levaram a criação das variáveis relacionadas a histórico pessoal de doenças, pouco contribuiu para a predição e isso pode se dar, em parte, pelo tamanho da amostra e prevalência reduzida destas doenças (CRAWFORD et al., 2005).

Quanto as variáveis relacionadas a prontidão para mudança, não foi possível identificar as suas contribuição para o modelo preditivo. Considerar as preferências e possibilidades dos pacientes é importante para o sucesso de programas de saúde e, por isso, uma vez identificados estes pacientes de maneira preditiva, considerar estas informações antes de efetivamente incluí-los em programas e linhas de cuidado, pode ser um fator crítico de sucesso destas ações, a fim de estabelecer expectativas de resultados e participação efetiva destes pacientes (KHULLAR; KAUSHAL, 2018; SMITH et al., 2019) e pode ser um critério de avaliação antes da efetiva adesão de pacientes.

Pacientes de alto custo muitas vezes recebem cuidados inapropriados para o tratamento de suas condições de saúde (SIEKMAN & HILGER, 2018) e o modelo proposto nesta pesquisa deixa evidente que diversas variáveis estão relacionadas a estes pacientes, ainda que não diretamente responsáveis pelo seu alto custo. Conhecê-las e considerá-las para a elaboração de programas de saúde poderá permitir uma gestão dos recursos otimizada e uma maior assertividade nas ações propostas, considerando outras condições que não somente a primeira causa da incorrência do alto custo. Por este motivo, modelos com resultados interpretáveis foram a preferência dentre os especialistas.

Esta unanimidade corrobora com o encontrado na literatura referente a predição de desfechos em saúde. McKelvey e colaboradores (2018), em artigo publicado sobre a interpretabilidade do aprendizado de máquina em saúde, destacam:

> Modelos interpretáveis de aprendizado de máquina permitem que os usuários finais avaliem o modelo, de preferência antes que uma ação seja realizada pelo usuário final, como por exemplo, o médico. Explicando o raciocínio por trás das previsões, sistemas de aprendizado de máquina interpretáveis fornecem aos usuários razões para aceitar ou rejeitar previsões e recomendações (MCKELVEY et al., 2018, p. 1).

Os algoritmos interpretáveis permitem um melhor uso prático das informações, em especial quando se trata de gestão de saúde e manejo de doenças, concluem os autores que utilizaram o algoritmo *random forest* para predição utilizando dados genômicos (CHEN; ISHWARAN, 2012).

Embora esta preferência exista, já vem sendo amplamente discutido na literatura o potencial do uso de algoritmos não interpretáveis na área da saúde, inclusive, com técnica para validação dos seus resultados, que envolve: a garantia da qualidade dos dados, testar o desempenho do algoritmo contra dados de testes independentes e avaliar seu desempenho em uso contínuo, com dados do mundo real (PRICE, 2018). Além da interpretabilidade, aspectos éticos também têm gerado discussões a respeito do uso destes algoritmos na área médica, em especial, quando envolvem o atendimento direto ao paciente. Usuários de modelos de aprendizado de máquina na área de saúde devem ter acesso não somente a um modelo preciso, mas também devem confiar na precisão do modelo e entender como o modelo funciona, qual sua recomendação e o porquê, para que possam propor e realizar ações em sua prática diária (YOON; TORRANCE; SCHEINERMAN, 2021).

6.2 LIMITAÇÕES

Dentre as limitações desta pesquisa, pode-se destacar, primeiramente, aquelas relacionadas aos questionários de saúde. No questionário utilizado, todos as questões eram de resposta obrigatória, evitando o viés de não resposta que, por um lado, facilitou o preparo do conjunto de dados, mas, por outro, pode conter nas respostas algo diverso da realidade de cada paciente.

As variáveis extraídas do questionário de saúde desta pesquisa retratam a avaliação e autopercepção dos pacientes em relação ao seu histórico de saúde e não necessariamente um diagnóstico feito por um profissional habilitado. Declarar-se deprimido, ansioso ou com transtorno bipolar, não necessariamente significa um diagnóstico confirmado por profissional médico. A confirmação das informações relacionadas a diagnósticos pode facilitar o planejamento e gerenciamento destes pacientes, uma vez identificados como alto custo pelo modelo ora proposto.

As limitações relacionadas aos vieses dos questionários de saúde como fontes de dados para predição são relatadas em estudos na literatura (BOSCARDIN et al., 2015; JOHNSTON; PROPPER; SHIELDS, 2009). Contudo, questionários ainda consistem em uma maneira fácil e de baixo custo para coletar dados de pacientes (BOSCARDIN et al., 2015).

Devido ao tamanho da amostra, não foi possível segmentar os pacientes pelas patologias autorreferidas e viabilizar modelos de predição de alto custo considerando doenças ou condições de saúde específicas, conforme o trabalho de Luo e colaboradores (2020) para predição de custo de pacientes com doença pulmonar obstrutiva crônica, e Wammes e colaboradores (2019) para predição de custo em pacientes com insuficiência cardíaca. A proposta separada por patologia poderia auxiliar no desenvolvimento de ações de prevenção e gerenciamento de pacientes em programas de manejo específico, como aqueles programas para pacientes diabéticos ou hipertensos. Em uma pesquisa que teve como objetivo avaliar a acurácia de redes neurais em predizer pacientes com alto custo, comparando amostras utilizando a população geral e segmentações de pacientes com asma, problemas cardíacos e diabetes, concluiu que a eficácia da predição varia substancialmente de acordo com a doença e que estes resultados podem estar associados ao tamanho das amostras (CRAWFORD et al., 2005).

É importante destacar, também, o cenário desta pesquisa: foram utilizados dados de 553 pacientes elegíveis ao plano de saúde por serem trabalhadores de uma empresa, o que torna a população desta pesquisa bastante específica a um determinado perfil, de pacientes em idade produtiva, embora os achados relacionados às características dos pacientes com alto custo estejam relacionadas aos achados em outros estudos (DOVE; DUNCAN; ROBB, 2003; ZULMAN et al., 2015). Considerando que a maior parte da população assistida por plano de saúde é elegível a cobertura pelo vínculo empregatício, este modelo pode ser replicado em outros cenários que

não somente em empresas, respeitando as características da população da pesquisa e os resultados explicados pelo algoritmo.

6.3 TRABALHOS FUTUROS

A partir dos resultados obtidos por esta pesquisa, é possível elencar propostas de trabalhos futuros.

a) Avaliar a evolução temporal da demanda por procedimentos de saúde e em que momento há este aumento de procedimentos que fazem com que o paciente se torne um alto custo, sendo possível criar alertas deste aumento para intervenções de saúde;
b) Avaliar o impacto nos custos laborais dos pacientes de alto custo, com a adição de dados de absenteísmo e afastamento;
c) Propor um modelo de identificação precoce de pacientes com transtornos mentais, sendo esta condição associada ao alto custo assistencial e afastamentos previdenciários, ambos com alto custo social;
d) Testar o modelo com amostras maiores e avaliar sua generalização ou necessidade de modelos específicos para a realidade de cada empresa (cenários diversos ainda que com características similares);
e) Identificar, por meio de algoritmos de regras de associação, como as variáveis selecionadas para esta pesquisa estão associadas;
f) Avaliar o ruído nos dados dos atributos utilizados nesta pesquisa e o seu possível impacto nos resultados;

7 CONSIDERAÇÕES FINAIS

Nesta pesquisa ficou evidenciado o potencial do uso de um algoritmo de aprendizado de máquina para a predição de pacientes de alto custo, com a utilização de dados administrativos e autorreferidos. Estes, não são comumente empregados em modelos preditivos e demonstram potencial preditivo e, mais importante, confirmam que com a explicação dos resultados do algoritmo, é possível identificar variáveis relacionadas a saúde que facilmente podem ser consideradas para implantação de programas de gestão e monitoramento.

O fato de as variáveis relacionadas a saúde mental terem grau de importância maior do que aquelas relacionadas a outras doenças crônicas pode, sim, estar relacionado ao tamanho e perfil da amostra e momento em que o questionário de saúde foi aplicado. Entretanto, foi consenso entre os especialistas que estas variáveis estão relacionadas ao desfecho alto custo, ainda que, pelo modelo, não seja possível identificar suas correlações com outras variáveis ou o motivo da incorrência do alto custo.

Considerar o cenário e a população estudada é importante para a interpretação dos resultados e para que ações de saúde personalizadas possam ser implantadas. As pesquisas recentes focadas em predição de pacientes de alto custo corroboram com a necessidade de modelos que considerem estes fatores, dada a heterogeneidade destes pacientes e de suas necessidades de saúde.

A saúde mental ganhou destaque nas empresas durante o período da pandemia e esta pesquisa contribui para evidenciar que a saúde mental dos colaboradores tem impacto em custos com saúde. Ainda, os afastamentos do trabalho gerados por doenças mentais geram prejuízos para o Estado, para famílias e para a economia, além daqueles custos com o plano de saúde, sendo um tema de relevância para a gestão de saúde neste momento e no futuro.

REFERÊNCIAS

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS). Portal da Agência Nacional de Saúde Suplementar. **Dados e indicadores do setor**. Disponível em: <http://www.ans.gov.br/perfil-do-setor/dados-e-indicadores-do-setor>. Acesso em: 15 dez. 2019.

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS). Portal da Agência Nacional de Saúde Suplementar. **Atualização do rol de procedimentos**. Disponível em: <http://www.ans.gov.br/participacao-da-sociedade/atualizacao-do-rol-de-procedimentos/como-e-atualizado-o-rol-de-procedimentos>. Acesso em: 10 fev. 2021a.

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR (ANS). Portal da Agência Nacional de Saúde Suplementar. **Tabela TUSS**. Disponível em: <http://www.ans.gov.br/images/stories/Legislacao/in/anexo_in34_dides.pdf>. Acesso em: 10 dez. 2021b.

ALBERGA, A.; HOLDER, L.; KORNAS, K.; BORNBAUM, C.; ROSELLA, L. Effects of behavioural risk factors on high-cost users of healthcare: a population-based study. **Canadian Journal of Public Health**, v. 109, n. 4, p. 441-450, 2018.

ALEXANDRE, N. M. C.; COLUCI, M. Z. O. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. **Ciência & Saúde Coletiva**, v. 16, n. 7, p. 3061-3068, jul. 2011.

ALZUBI, J.; NAYYAR, A.; KUMAR, A. Machine Learning from Theory to Algorithms: An Overview. **Journal of Physics Conference Series**, v. 1142, n. 1, 2018.

ANDRADE, A. **Best practices for convolutional neural networks applied to object recognition in images**. Toronto: University of Toronto, 2013.

ARAÚJO, A. A. S.; SILVA, J. R. S. Análise de tendência da sinistralidade e impacto na diminuição do número de operadoras de saúde suplementar no Brasil. **Ciência & Saúde Coletiva**, v. 23, n. 8, p. 2763-2770, 2018.

BATES, D. W.; SARIA, S.; OHNO-MACHADO, L.; SHAH, A.; ESCOBAR, G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. **Health Affairs**, v. 33 n. 7, p. 1123-1131, 2014.

BEAM, A.; KOHANE, I. Big Data and Machine Learning in Health Care. **JAMA**, v. 319, n. 13, p. 1317-1318, 2018.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, n. 1, p.281-305, 2012.

BERTSIMAS, D.; BJARNADÓTTIR, M.; KANE, M.; KRYDER, J.; PANDEY, R.; VEMPALA, S.; WANG, G. Algorithmic Prediction of Health Care Costs and Discovery of Medical Knowledge. **Operations Research**, v. 56, n.6, p. 1382-1392, 2008.

BIAU, G.; SCORNET, E. A random forest guided tour. **TEST**, v. 25, p. 197-227, 2016.

BLUMENTHAL, D.; CHERNOF, B.; FULMER, T.; LUMPKIN, J.; SELBERG, J. Caring for High-Need, High-Cost Patients - An Urgent Priority. **New England Journal of Medicine**, v. 375, n. 10, p. 909-911, 2016.

BOSCARDIN, C. K.; GONZALES, R.; BRADLEY, K. L.; RAVEN, M. C. Predicting cost of care using self-reported health status data. **BMC Health Services Research**, v. 15, n. 1, p. 406, 2015.

BOSE, E.; RADHAKRISHNAN, K. Using Unsupervised Machine Learning to Identify Subgroups Among Home Health Patients with Heart Failure Using Telehealth. **Computers Informatics Nursing**, v. 36, n. 5, p. 242-248, 2018.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123-140, 1996.

_____. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BUCK, J. A.; TEICH, J. L.; MILLER, K. Use of Mental Health and Substance Abuse Services Among High-Cost Medicaid Enrollees. **Administration and Policy in Mental Health**, v. 31, n. 1, p. 3-14 2003.

CAETANO, R.; SILVA, A. B.; GUEDES, A. C. C. M.; PAIVA, C. C. N.; RIBEIRO, G. R.; SANTOS, D. L.; SILVA, R. M. Desafios e oportunidades para telessaúde em tempos da pandemia pela COVID-19: uma reflexão sobre os espaços e iniciativas no contexto brasileiro. **Cadernos de Saúde Pública**, v. 36, n. 5, 2020.

CARVALHO, D. R.; DALAGASSA, M. R; SILVA, S. H. Uso de técnicas de mineração de dados para a identificação automática de beneficiários propensos ao diabetes mellitus tipo 2. **Informação & Informação**, v. 20, n. 3, p. 274-296, 2015.

CASTRO, M. M. C.; QUARANTINI, L. C.; DALTRO, C.; PIRES-CALDAS, M.; KOENEN, K. C.; KRAYCHETE, D. C.; de OLIVEIRA, I. R. Comorbidade de sintomas ansiosos e depressivos em pacientes com dor crônica e o impacto sobre a qualidade de vida. **Revista de Psiquiatria Clínica**, São Paulo, v. 38, n. 4, p. 126-129, 2011.

CHECHULIN, Y.; NAZERIAN, A.; RAIS, S.; MALIKOV, K. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). **Healthcare Policy**, v. 9, n. 3, p. 68-79, 2014.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, v. 99, n. 6, p. 323-329, , 2012.

CHOLLET, F. **Deep learning with Python**. [S.l.]: Manning Publications, 2018.

COUGHLIN, T. A.; LONG, S. K. Health care spending and service use among high-cost Medicaid beneficiaries, 2002-2004. **Inquiry**, v. 4, n. 4, p. 405-417, 2009.

CRAWFORD, A.; FUHR, J.; CLARKE, J.; HUBBS, B. Comparative Effectiveness of Total Population versus Disease-Specific Neural Network Models in Predicting Medical Costs. **Disease Management: DM**, v. 8, n. 5, p. 277-287, 2005.

CUNNINGHAM, P. J. Predicting high-cost privately insured patients based on self-reported health and utilization data. **American Journal of Managed Care**, v. 23, n. 7, p. 215-222, 2017.

DE OLIVEIRA, C.; PATAKY, R.; BREMNER, K. E.; RANGREJ , J.; CHAN, K. K. W.; CHEUNG, W. Y.; HOCH, J. S.; PEACOCK, S.; KRAHN, M. D. Estimating the cost of cancer care in British Columbia and Ontario: a Canadian inter-provincial comparison. **Healthc Policy**, v. 12, n. 3, p. 95-108, 2017.

DeSALVO, K. B.; JONES, T. M.; PEABODY, J.; MCDONALD, J.; FIHN, S.; FAN, V.; HE, J.; MUNTNER, P. Health care expenditure prediction with a single item, self-rated health measure. **Medical Care**, v. 47, n. 4, p. 440-447, 2009.

DIEHR, P.; YANEZ, D.; ASH, A.; HORNBROOK, M.; LIN, D. Methods for analyzing healthcare utilization and costs. **Annual Review Of Public Health**, v. 20, n.1, p. 125-144, 1999.

DOUPE, P.; FAGHMOUS, J.; BASU, S. Machine Learning for Health Services Researchers. **Value in Health**, v. 22, n. 7, p. 808-815, 2019.

DOVE, H.; DUNCAN, I.; ROBB, A. A prediction model for targeting low-cost, high-risk members of managed care organizations. **American Journal of Managed Care**, v. 9, n. 5, p. 381-389, 2003.

DUARTE, A. L. C. M.; OLIVEIRA, F. M.; SANTOS, A. A.; SANTOS, B. F. C. Evolução na utilização e nos gastos de uma operadora de saúde. **Ciência e Saúde Coletiva**, v. 22, n. 8, p. 2753-2759, 2017.

DUNCAN, I.; LOGINOV, M.; LUDKOVSKI, M. Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. **North American Actuarial Journal**, v. 20, n. 1, p. 65-87, 2016.

DYNIEWICZ, A. M. **Metodologia da pesquisa em saúde para iniciantes**. 3. ed. São Caetano do Sul: Difusão, 2014.

FERVER, K.; BURTON, B.; JESILOW, P. The Use of Claims Data in Healthcare Research. **Open Public Health Journal**, v. 2, p. 11-24, 2009.

FIGUEROA, J. F.; ZHOU, X.; JHA, A. K. Characteristics and Spending Patterns of Persistently High-Cost Medicare Patients. **Health Affairs**, v. 38, n.1, p.107-114, 2019.

FILGUEIRAS, A.; GALVÃO, B. O.; PIRES, P.; FIORAVANTI-BASTOS, A. C. M.; HORA, G. P. R.; SANTANA, C. M. T.; LANDEIRA-FERNANDEZ, J. Translation and semantic adaptation of the attentional control scale for the Brazilian context. Tradução e adaptação semântica do questionário de controle atencional para o contexto brasileiro. **Estudos de Psicologia**, Campinas, v. 32, n. 2, p. 173-186, 2015.

FLEISHMAN ,J. A.; COHEN J. Using information on clinical conditions to predict high-cost patients. **Health Services Research**, v. 45, n. 2, p. 532-552, 2010.

GASKIN, D. J.; RICHARD, P. The Economic Costs of Pain in the United States. In: INSTITUTE OF MEDICINE (US) COMMITTEE ON ADVANCING PAIN RESEARCH, CARE, AND EDUCATION. **Relieving Pain in America**: A Blueprint for Transforming Prevention, Care, Education, and Research. Washington (DC): National Academies Press (US), 2011. Appendix C. Disponível em: <www.ncbi.nlm.nih.gov/books/NBK92521/> Acesso em 16 dez. 2021.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **Weka**: the Weka data mining software. 2009. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 25 jan. 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. New York: Springer; 2008.

HAYKIN, S. **Redes neurais**: princípios e práticas. Tradução de P. M. Engel. 2. ed. Porto Alegre: Bookman, 2001.

HERNÁNDEZ NIETO, H. **Contributions to statistical analysis**: The Coefficients of Proportional Variance, Content Validity and Kappa. Merida (Venezuela): BookSurge Publishing, 2002.

HIBBARD, J. H.; GREENE, J.; SACKS, R.; OVERTON, V.; PARROTTA, C. D. Adding A Measure of Patient Self-Management Capability to Risk Assessment Can Improve Prediction of High Costs. **Health Affairs**, v. 35 n. 3, p. 489-494, 2016.

HOGAN, M. E.; TADDIO, A.; KATZ, J.; SHAH, V.; KRAHN, M. Incremental health care costs for chronic pain in Ontario, Canada: a population-based matched cohort study of adolescents and adults using administrative data. **Pain**, v. 157, n. 8, p. 1626-1633, 2016.

INSTITUTO DE ESTUDOS DE SAÚDE SUPLEMENTAR (IESS). **Caracterização dos beneficiários com alto custo assistencial**: um estudo de caso. 2017. Disponível em: <https://www.iess.org.br/?p=publicacoes&id=878&id_tipo=15>. Acesso em: 14 dez. 2020.

INSTITUTO DE ESTUDOS DE SAÚDE SUPLEMENTAR (IESS). **Despesas assistenciais das operadoras**. Disponível em: <https://www.iess.org.br/?p=blog&id=943>. Acesso em: 10 nov. 2020.

INSTITUTO DE ESTUDOS DE SAÚDE SUPLEMENTAR (IESS). **Como está a gestão da saúde entre as empresas**. Disponível em: <https://iess.org.br/?p=blog&id=1262>. Acesso em: 10 fev. 2021a.

INSTITUTO DE ESTUDOS DE SAÚDE SUPLEMENTAR (IESS). IESS Data. **Total de beneficiários de planos de assistência médico hospitalar**. Disponível em: <https://iessdata.iess.org.br/home>. Acesso em: 22 nov. 2021b.

JOHNSTON, D. W.; PROPPER, C.; SHIELDS, M. A. Comparing subjective and objective measures of health: Evidence from hypertension for the income/health gradient. **Journal of Health Economics**, v. 28, n. 3, p. 540-552, 2009.

JOYNT, K.; GAWANDE, A. A.; ORAV, E.; JHA, A. K. Contribution of preventable acute care spending to total spending for high-cost Medicare patients. **JAMA**, v. 309, n. 24, p. 2572-2578, 2013.

KHULLAR, D.; KAUSHAL, R. "Precision health" for high-need, high-cost patients. **American Journal of Managed Care**, v. 24. n. 9, p. 396-398, 2018.

KIM, M. H.; BANERJEE, S.; PARK, S. M.; PATHAK, J. Improving risk prediction for depression via Elastic Net regression - Results from Korea National Health Insurance Services Data. **AMIA Annual Symposium Proceedings**, v. 2016, p.1860-1869, 2017.

KIM, Y.; PARK, H. Improving Prediction of High-Cost Health Care Users with Medical Check-Up Data. **Big Data**. v. 7, n. 3, p.163-175, 2019.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York: Springer; 2013.

LEE, J. Y.; MURATOV, S.; TARRIDE, J.-E.; HOLBROOK, A. M. Managing high-cost healthcare users: the international search for effective evidence-supported strategies. **Journal of the American Geriatrics Society**, v. 66, n. 5, p. 1002-1008, 2018.

LEE, Y. C.; HUANG, S. C.; HUANG, C. H.; WU, H. H. A new approach to identify high burnout medical staffs by kernel K-means cluster analysis in a regional teaching hospital in Taiwan. **Inquiry**, v. 53, n. 5, 2016.

LENTZ, T. A.; HARMAN, J. S.; MARLOW, N. M.; BENECIUK, J. M.; FILLINGIM, R. B.; GEORGE, S. Z. Factors associated with persistently high-cost health care utilization for musculoskeletal pain. **PLoS One**, v. 14, n. 11, p. e0225125, 2019.

LIKERT, R. A technique for the measurement of attitudes. **Archives of Psychology**, v. 22, n. 140, p. 55, 1932.

LOPES, B.; RAMOS, I. C. O.; RIBEIRO, G.; CORREA, R.; VALBON, B. F.; LUZ, A. C.; SALOMÃO, M.; LYRA, J. M.; AMBRÓSIO JUNIOR, R. Bioestatísticas: conceitos fundamentais e aplicações práticas. **Revista Brasileira de Oftalmologia**, Rio de Janeiro, v. 73, n. 1, p. 16-22, 2014.

LUO, L.; LI, J.; LIAN, S.; ZENG, X.; SUN, L.; LI, C.; HUANG, D.; ZHANG, W. Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China. **Health Informatics Journal**, v. 26, n.3, p. 1577-1598, 2020.

LYNN, M. R. Determination and quantification of content validity. **Nursing Research**, v. 35, n. 6, p. 382-385, 1986.

MARSLAND, S. **Machine learning**: an algorithmic perspective. Boca Raton: CRC, 2015.

MCKELVEY, T.; AHMAD, Muhammad; TEREDESAI, Ankur; ECKERT, Carly. Interpretable Machine Learning in Healthcare. **IEEE Intelligent Informatics Bulletin**, v. 19, n. 1, 2018.

MENDES, E. A.; TEIXEIRA, L.; BONFATTI, R. J. As condições de saúde dos trabalhadores a partir dos exames periódicos de saúde. **Saúde Debate**, Rio de Janeiro, v. 41, n. 112, p. 142-154, jan./mar. 2017.

MENDES, E. V. **O cuidado das condições crônicas na atenção primária à saúde**: o imperativo da consolidação da estratégia da saúde da família. Brasília: Organização Pan-Americana da Saúde, 2012.

MOHNEN, S. M.; ROTTEVEEL, A. H.; DOORNBOS, G.; POLDER, J. J. Healthcare Expenditure Prediction with Neighborhood Variables – A Random Forest Model, Statistics. **Politics and Policy**, v. 11, n.2, p. 111-138, 2020.

MORID, M. A.; KAWAMOTO, K.; AULT, T.; DORIUS, J.; ABDELRAHMAN, S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. **AMIA Annual Symposium Proceedings**, v. 2017, p. 1312-1321, 2018.

MOTURU, S. T; LIU, H.; JOHNSON, W. G. Healthcare risk modeling for medicaid patients: the impact of sampling on the prediction of high-cost patients. In: HEALTHINF 2008 - INTERNATIONAL CONFERENCE ON HEALTH INFORMATICS, 1., 2008, Madeira. **Proceedings**..., Madeira, Portugal, 2008. p. 126-133.

MREJENS, M.; RACHE, B.; NUNES, L. COVID-19 e Saúde Mental: Uma Análise de Tendências Recentes no Brasil. **Instituto de Estudos para Políticas de Saúde**, maio, 2021. (Nota Técnica 20). Disponível em: < https://ieps.org.br/pesquisas/covid-19-e-saude-mental-uma-analise-de-tendencias-recentes-no-brasil/ >. Acesso em: 17 jan. 2022.

NADERIFAR, M.; GOLI, H.; GHALJAIE, F. Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. **Strides in Development of Medical Education**, v. 14, n. 3, p. e67670, 2017.

NEMBRINI, S.; KÖNIG, I.; WRIGHT, M. The revival of the Gini Importance? **Bioinformatics**, v. 34, n. 21, p. 3711-3718, 2018.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Medication Without Harm**. Disponível em: <https://www.who.int/initiatives/medication-without-harm>. Acesso em: 14 dez. 2021.

OSAWA, I.; GOTO, T.; YAMAMOTO, Y.; TSUGAWA, Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. **NPJ Digital Medicine**, v. 3, n. 1, p. 148, 2020.

PANAY, B.; BALOIAN, N.; PINO, J. A.; PEÑAFIEL, S.; SANSON, H.; BERSANO, N. Predicting Health Care Costs Using Evidence Regression. **Proceedings**, v. 31, n. 1, p. 74. 2019.

_____. Feature Selection for Health Care Costs Prediction Using Weighted Evidential Regression. **Sensors**, v. 20, n. 16, p. 4392, 2020.

PERRIN, N. A.; STIEFEL, M.; MOSEN, D. M.; BAUCK, A.; SHUSTER, E.; DIRKS, E. M. Self-reported health and functional status information improves prediction of inpatient admissions and costs. **American Journal of Managed Care**, v. 17, n. 12, p. 472-478, 2011.

PIRES, F. M. S. **Estudo do impacto da medicina preventiva na diminuição da sinistralidade dos planos de saúde e sua aplicação ao sistema SAMMED/FUSEX**. Rio de Janeiro: Escola de Saúde do Exército, 2008.

PRICE, W. N. Big data and black-box medical algorithms. **Science Translational Medicine**, v. 10, n. 471, p. eaao5333, 2018.

RODRIGUES, J. A. R. M.; CUNHA, I. C. K. O.; VANNUCHI, M. T. O.; HADDAD, M. C. F. L. Out-of-pocket payments in hospital bills: a challenge to management. **Revista Brasileira de Enfermagem**, v. 71, n. 5, p. 2511-2518, 2018.

ROSAS, M.; BEZERRA, A.; DUARTE-NETO, P. Uso das redes neurais artificiais na aplicação de metodologia para alocação de recursos da saúde. **Revista de Saúde Pública**, v. 47, p.128-136, 2013.

ROSELLA, L. C.; FITZPATRICK, T.; WODCHIS, W. P.; CALZAVARA, A.; MANSON, H.; GOEL, V. High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. **BMC Health Services Research**, v. 14, n. 532, 2014.

SALISBURY, C.; JOHNSON, L.; PURDY, S.; VALDERAS, J. M.; MONTGOMERY, A. A. Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study. **British Journal of General Practice**, v. 61, n. 582, p. 12-21, 2011.

SAMUEL, L. A. Some studies in machine learning using the game of checkers. **IBM Journal**, v. 3, n. 3, p. 210-229, 1959.

SANTOS, A. M.; SEIXAS, J. M.; PEREIRA, B. B.; MEDRONHO, R. A. Usando redes neurais artificiais e regressão logística na predição da Hepatite A. **Revista Brasileira de Epidemiologia**, v. *8*, n. 2, p. 117-126, 2005.

SANTOS, H. G. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina**. Tese (Doutorado) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2018.

SARICA, A.; CERASA, A.; QUATTRONE, A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. **Frontiers in Aging Neuroscience**, v. 9, p. 329, 2017.

SEDGWICK, P. Snowball sampling. **BMJ**, v. 347, n. 2, p. 1-2, 2013.

SHAHID, N.; RAPPON, T.; BERTA, W. Applications of artificial neural networks in health care organizational decision-making: a scoping review. **PLoS One**, v. 14, n. 2, 2019.

SHERMAN, B. W.; FABIUS, R. J. Quantifying the value of worksite clinic nonoccupational health care services: a critical analysis and review of the literature. **Journal of Occupational and Environmental Medicine**, v. 54, n. 4, p. 394-403, 2012.

SHRESTHA, A.; BERGQUIST, S.; MONTZ, E.; ROSE, S. Mental Health Risk Adjustment with Clinical Categories and Machine Learning. **Health Services Research**, v. 53, Suppl. 1, p. 3189-3206, 2018.

SIDEY-GIBBONS, J.; SIDEY-GIBBONS, C. Machine learning in medicine: a practical introduction. **BMC Medical Research Methodology**, v. 19, n. 1, p. 64, 2019.

SIEKMAN, N.; HILGER, R. High users of healthcare: Strategies to improve care, reduce costs. **Cleveland Clinic Journal of Medicine**, v. 85, n.1, p. 25-31, 2018.

SILVA, L. E.; COHEN, R. V.; ROCHA, J. L. L.; ROCHA, J. L. L.; HASSEL, V. M. C. Cirurgias eletivas no novo normal pós-pandemia da COVID-19: testar ou não testar? **Revista do Colégio Brasileiro de Cirurgiões**, v. 47, p. e20202649, 2020.

SMEETS, R. G. M.; ELISSEN, A. M. J.; KROESE, M. E. A. L.; HAMELEERS, N.; RUWAARD, D. Identifying subgroups of high-need, high-cost, chronically ill patients in primary care: A latent class analysis. **PLoS One**, v. 15, n. 1, p. e0228103, 2020.

SMITH, T. G.; DUNN, M. E.; LEVIN, K. Y.; TSAKRAKLIDES, S. P.; MITCHELL, S. A.; van de POLL-FRANSE, L. V.; WARD, K. C.; WIGGINS, C. L.; WU, X. C.; HURLBERT, M.; AARONSON, N. K. Cancer survivor perspectives on sharing patient-generated health data with central cancer registries. **Quality of Life Research**, v. 28.n. 11, p. 2957-2967, 2019.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing and Management**, n. 45, p. 427-437, 2009.

SUSHIMITA, S.; NEWMAN, S.; MARQUARDT, J.; RAM, P.; PRASAD, V.; De COCK, M.; TEREDESAI, A. Population Cost Prediction on Public Health Datasets. **ACM International Conference Proceeding Series**, 2015. p. 87-94.

TANKE, M. A. C.; FEYMAN, Y.; BERNAL-DELGADO, E.; DEENY, S. R.; IMANAKA, Y.; JEURISSEN, P.; LANGE, L.; PIMPERL, A.; SASAKI, N.; SCHULL, M.; WAMMES, J. J. G.; WODCHIS, W. P.; MEYER, G. S. A challenge to all. A primer on inter-country differences of high-need, high-cost patients. **PLoS One**, v. 14, n. 6, p. e0217353, 2019.

TAVARES, R. S. C.; KAMIMURA, Q. P. Saúde Corporativa:um olhar estratégico para o capital humano em indústria do setor automobilístico localizado no Vale do Paraíba. **Revista Científica On-line Tecnologia Gestão Humanismo**, v. 4, p. 72-83, 2014.

THEOBALD, O. **Machine Learning for Absolute Beginners**: A Plain English Introduction. [S.l.]: Scatterplot Press, 2017.

VIEIRA, T. W.; SAKAMOTO, V. T. M.; de MORAES, L. C.; BLATT, C. R.; CAREGNATO, R. C. A. Validation methods of nursing protocols: an integrative review. **Revista Brasileira de Enfermagem**, v. 73, n. Suppl. 5, p. 1-10, 2020.

WAMMES, J. J. G.; AUENER, S.; VAN DER WEES, P. J.; TANKE , M. A. C.; BELLERSEN, L.; WESTERT, G. P.; ATSMA, F.; JEURISSEN, P. P. T. Characteristics and health care utilization among patients with chronic heart failure: a longitudinal claim database analysis. **ESC Heart Fail**, v. 6, n. 6, p. 1243-1251, 2019.

WAMMES, J. J. G.; TANKE, M.; JONKERS, W.; WESTERT, G. P.; P van der WEES, J.; JEURISSEN, P. P. T. Characteristics and healthcare utilization patterns of high-cost beneficiaries in the Netherlands: a cross-sectional claims database study. **BMJ Open**. v. 7, n. 11, p. e017775, 2017.

WAMMES, J. J. G.; VAN DER WEES, P. J.; TANKE, M. A. C.; WESTERT, G. P.; JEURISSEN, P. P. T. Systematic review of high-cost patients' characteristics and healthcare utilisation. **BMJ Open**, v. 8, n. 9, p. e023113, 2018.

WERE, M. C.; KAMANO, J. H.; VEDANTHAN, R. Leveraging Digital Health for Global Chronic Diseases. **Global Heart**, v. 11, n.4, p. 459-462, 2016.

WEYMANN, D.; SMOLINA, K.; GLADSTONE, E. J.; MORGAN, S. G. High-Cost Users of Prescription Drugs: A Population-Based Analysis from British Columbia, Canada. **Health Services Research**, v. 52, n. 2, p. 697-719, 2017.

WODCHIS, W. P.; AUSTIN, P. C.; HENRY, D. A. A 3-year study of high-cost users of health care. **CMAJ**, v. 188, n. 3, p. 182-188, 2016.

WORLD HEALTH ORGANIZATION. Obesity: preventing and managing the global epidemic. **WHO Technical Report Series**, n. 894, p. 1-253, 2000.

YANG, C.; DELCHER, C.; SHENKMAN, E.; RANKA, S. Machine learning approaches for predicting high-cost high need patient expenditures in health care. **BioMedical Engineering OnLine**, v. 17, Suppl. 1, p. 131, 2018.

YANG, F.; WANG, H.; MI, H.; LIN, C.-de; CAI, W.-wen. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. **BMC Bioinformatics**, v. 10, Suppl. 1, p. S22, 2009.

YOON, C. H.; TORRANCE, R.; SCHEINERMAN, N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? **Journal of Medical Ethics**, p. 1-5, 2021.

ZHANG, C.; CAO, L.; ROMAGNOLI, A. On the feature engineering of building energy data mining. **Sustainable Cities and Society**, v. 39, p. 508-518, 2018.

ZIROLDO, R. R.; GIMENES, R. O.; CASTELO JÚNIOR, C. A importância da saúde suplementar na demanda da prestação dos serviços assistenciais no Brasil. **O Mundo da Saúde**, São Paulo, v. 37, n. 2, p. 216-221, 2013.

ZULMAN, D. M.; CHEE, C. P. C.; WAGNER, T. H.; YOON, J.; COHEN, D. M.; HOLMES, T. H.; RITCHIE, C.; ASCH, S. M. Multimorbidity and healthcare utilization among high-cost patients in the US Veterans Affairs Health Care System. **BMJ Open**, v. 5, n. 4: e007771, 2015.

APÊNDICE A

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Você está sendo convidado(a) como voluntário(a) a participar do estudo “Identificação de Pacientes de Alto Custo com a Utilização de Dados Administrativos e Autorreferidos por Meio do Aprendizado de Máquinas”, que tem como objetivo identificar pacientes de alto custo a partir de dados administrativos, comportamentais e socioeconômicos. Este estudo é importante porque os pacientes com alto custo consomem grande parte dos recursos dos sistemas de saúde e muitas vezes recebem um cuidado inapropriado para suas necessidades. Com um modelo de identificação precoce destes pacientes, é possível o melhor gerenciamento destes recursos e das condições de saúde destes pacientes, indicando o melhor tratamento e possibilitando a prevenção de desfechos de saúde indesejados e custosos.

PARTICIPAÇÃO NO ESTUDO

A sua participação compreende analisar os resultados obtidos e a respectiva contribuição das variáveis. Após avaliar os resultados, você deverá assinalar, de acordo com a concordância: concordo, concordo parcialmente e discordo. Estima-se que serão necessários 10 minutos para o preenchimento deste questionário. O local para preenchimento é de livre escolha.

RISCOS E BENEFÍCIOS

Por meio deste Termo de Consentimento Livre e Esclarecido você está sendo alertado de que não existe benefício direto. Porém, sua participação permite receber informações atualizadas sobre identificação precoce de pacientes com alto custo, além de vislumbrar o potencial do uso da inteligência artificial para a gestão de saúde. Em relação aos riscos, você poderá sofrer constrangimento, porém pode se opor a responder qualquer questão ou deixar de responder a qualquer momento. Alertamos que você pode se recusar a participar do estudo, ou retirar seu consentimento a qualquer momento, sem precisar justificar ou sofrer qualquer prejuízo.

Este estudo foi realizado com base de dados secundários, impossibilitando que tanto os pesquisadores quanto os especialistas identifiquem os pacientes, cujos dados foram analisados, minimizando quaisquer riscos de referências a dados sensíveis.

SIGILO E PRIVACIDADE

Nós pesquisadores garantiremos a você que sua privacidade será respeitada, ou seja, seu nome ou qualquer outro dado ou elemento que possa, de qualquer forma, lhe identificar, será mantido em sigilo. Nós pesquisadores nos responsabilizaremos pela guarda e confidencialidade dos dados, bem como a não exposição dos dados de pesquisa.

AUTONOMIA

Nós lhe asseguramos assistência durante toda pesquisa, bem como garantiremos seu livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que você queira saber antes, durante e depois de sua participação. Também informamos que você pode se recusar a participar do estudo, ou retirar seu consentimento a qualquer momento, sem precisar justificar, e de, por desejar sair da pesquisa, não sofrerá qualquer prejuízo à assistência que vem recebendo.

RESSARCIMENTO E INDENIZAÇÃO

Caso tenha qualquer despesa decorrente da participação nesta pesquisa, tais como transporte, alimentação entre outros, bem como de seu acompanhante, haverá ressarcimento dos valores gastos na forma seguinte: mediante depósito em conta corrente. De igual maneira, caso ocorra algum dano decorrente de sua participação no estudo, você será devidamente indenizado, conforme determina a lei.

CONTATO

Os pesquisadores envolvidos com o referido projeto são: Deborah Ribeiro Carvalho, Pontifícia Universidade Católica do Paraná – Programa de Pós-graduação em Tecnologia em Saúde e Ana Luísa Gonçalves Gomes Coelho Seleme, Pontifícia Universidade Católica do Paraná – Programa de Pós Graduação em Tecnologia em Saúde, com as quais você poderá manter contato pelo telefone (41) 99934-9080 ou (41) 99996-9394.

O Comitê de Ética em Pesquisa em Seres Humanos (CEP) é composto por um grupo de pessoas que estão trabalhando para garantir que seus direitos como participante de pesquisa sejam respeitados. Ele tem a obrigação de avaliar se a pesquisa foi planejada e se está sendo executada de forma ética. Se você achar que a pesquisa não está sendo realizada da forma como você imaginou ou que está sendo prejudicado de alguma forma, você pode entrar em contato com o Comitê de Ética em Pesquisa da PUCPR (CEP) pelo telefone (41) 3271-2103 entre segunda e sexta-feira das 08h00 às 17h30 ou pelo e-mail nep@pucpr.br.

DECLARAÇÂO

Declaro que li e entendi todas as informações presentes neste Termo de Consentimento Livre e Esclarecido e tive a oportunidade de discutir as informações deste termo. Todas as minhas perguntas foram respondidas e eu estou satisfeito com as respostas. Entendo que receberei uma via assinada e datada deste documento e que outra via assinada e datada será arquivada nos pelo pesquisador responsável do estudo.

Enfim, tendo sido orientado quanto ao teor de todo o aqui mencionado e compreendido a natureza e o objetivo do já referido estudo, manifesto meu livre consentimento em participar, estando totalmente ciente de que não há nenhum valor econômico, a receber ou a pagar, por minha participação.

Dados do participante da pesquisa	
Nome:	
Telefone:	
e-mail:	

Local, _____ de _____ de _____.

_____ _____

Assinatura do participante da pesquisa | Assinatura do Pesquisador

_____ _____

Assinatura do participante da pesquisa | Assinatura do Pesquisador

APÊNDICE B
INSTRUMENTO DE AVALIAÇÃO DOS ESPECIALISTAS

Instrumento de Avaliação dos Especialistas em Gestão de Saúde

Programa de Pós-graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná

Título da Pesquisa: Identificação de Pacientes de Alto Custo com a Utilização de Dados Administrativos e Autorreferidos por meio do Aprendizado de Máquina

Prezado Especialista:

Agradecemos por poder contar com sua contribuição nesta pesquisa que tem por objetivo Identificar pacientes de alto custo a partir de dados administrativos e autorreferidos por meio do aprendizado de máquina.

Sua participação é voluntária e sua identidade e as informações serão preservadas, conforme parecer de pesquisa aprovado no Comitê de Ética em Pesquisa PUCPR, sob o Número: 4.756.053, CAAE: 46628421.3.0000.0020, em 06 de junho de 2021.

Tempo estimado para resposta 10 minutos.

*Obrigatório

1. E-mail *

Termo de Consentimento Livre e Esclarecido (TCLE)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Você está sendo convidado(a) como voluntário(a) a participar do estudo "IDENTIFICAÇÃO DE PACIENTES DE ALTO CUSTO COM A UTILIZAÇÃO DE DADOS ADMINISTRATIVOS E AUTORREFERIDOS POR MEIO DO APRENDIZADO DE MÁQUINA", que tem como objetivo identificar pacientes de alto custo a partir de dados administrativos, comportamentais e socioeconômicos. Este estudo é importante porque os pacientes com alto custo consomem grande parte dos recursos dos sistemas de saúde e muitas vezes recebem um cuidado inapropriado para suas necessidades. Com um modelo de identificação precoce destes pacientes, é possível o melhor gerenciamento destes recursos e das condições de saúde destes pacientes, indicando o melhor tratamento e possibilitando a prevenção de desfechos de saúde indesejados e custosos.

PARTICIPAÇÃO NO ESTUDO

A sua participação compreende analisar os resultados obtidos e a respectiva contribuição das variáveis. Após avaliar os resultados, você deverá assinalar, de acordo com a concordância: concordo, concordo parcialmente e discordo. Estima-se que serão necessários 10 minutos para o preenchimento deste questionário. O local para preenchimento é de livre escolha.

RISCOS E BENEFÍCIOS

Por meio deste Termo de Consentimento Livre e Esclarecido você está sendo alertado de que não existe benefício direto. Porém, sua participação permite receber informações atualizadas sobre identificação precoce de pacientes com alto custo, além de vislumbrar o potencial do uso da inteligência artificial para a gestão de saúde. Em relação aos riscos, você poderá sofrer constrangimento, porém pode se opor a responder qualquer questão ou deixar de responder a qualquer momento. Alertamos que você pode se recusar a participar do estudo, ou retirar seu consentimento a qualquer momento, sem precisar justificar ou sofrer qualquer prejuízo.
Este estudo foi realizado com base de dados secundários, impossibilitando que tanto os pesquisadores quanto os especialistas identifiquem os pacientes, cujos dados foram anonimizados, minimizando quaisquer riscos de referências a dados sensíveis.

SIGILO E PRIVACIDADE

Nós pesquisadores garantiremos a você que sua privacidade será respeitada, ou seja, seu nome ou qualquer outro dado ou elemento que possa, de qualquer forma, lhe identificar, será mantido em sigilo. Nós pesquisadores nos responsabilizaremos pela guarda e confidencialidade dos dados, bem como a não exposição dos dados de pesquisa.

AUTONOMIA

Nós lhe asseguramos assistência durante toda pesquisa, bem como garantiremos seu livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que você queira saber antes, durante e depois de sua participação. Também informamos que você pode se recusar a participar do estudo, ou retirar seu consentimento a qualquer momento, sem precisar justificar, e de, por desejar sair da pesquisa, não sofrerá qualquer prejuízo à assistência que vem recebendo.

RESSARCIMENTO E INDENIZAÇÃO

Caso tenha qualquer despesa decorrente da participação nesta pesquisa, tais

como transporte, alimentação entre outros, bem como de seu acompanhante, haverá ressarcimento dos valores gastos na forma seguinte: mediante depósito em conta corrente. De igual maneira, caso ocorra algum dano decorrente de sua participação no estudo, você será devidamente indenizado, conforme determina a lei.

CONTATO

Os pesquisadores envolvidos com o referido projeto são: Deborah Ribeiro Carvalho, Pontifícia Universidade Católica do Paraná – Programa de Pós-graduação em Tecnologia em Saúde e Ana Luísa Gonçalves Gomes Coelho Seleme, Pontifícia Universidade Católica do Paraná – Programa de Pós Graduação em Tecnologia em Saúde, com as quais você poderá manter contato pelo telefone (41) 99934-9080.

O Comitê de Ética em Pesquisa em Seres Humanos (CEP) é composto por um grupo de pessoas que estão trabalhando para garantir que seus direitos como participante de pesquisa sejam respeitados. Ele tem a obrigação de avaliar se a pesquisa foi planejada e se está sendo executada de forma ética. Se você achar que a pesquisa não está sendo realizada da forma como você imaginou ou que está sendo prejudicado de alguma forma, você pode entrar em contato com o Comitê de Ética em Pesquisa da PUCPR (CEP) pelo telefone (41) 3271-2103 entre segunda e sexta-feira das 08h00 às 17h30 ou pelo e-mail nep@pucpr.br.

DECLARAÇÃO

Declaro que li e entendi todas as informações presentes neste Termo de Consentimento Livre e Esclarecido e tive a oportunidade de discutir as informações deste termo. Todas as minhas perguntas foram respondidas e eu estou satisfeito com as respostas. Entendo que receberei uma via assinada e datada deste documento e que outra via assinada e datada será arquivada nos pelo pesquisador responsável do estudo.

Enfim, tendo sido orientado quanto ao teor de todo o aqui mencionado e compreendido a natureza e o objetivo do já referido estudo, manifesto meu livre consentimento em participar, estando totalmente ciente de que não há nenhum valor econômico, a receber ou a pagar, por minha participação.

Concordância com o Termo de Consentimento Livre Esclarecido

2. Concorda com o TCLE? *

Marcar apenas uma oval.

- ◯ Concordo *Pular para a pergunta 4*
- ◯ Discordo

Endereço de E-mail (o endereço de e-mail é solicitado para o recebimento do formulário com as respostas tanto na concordância ou discordância em participar desta avaliação)

3. E-mail: *

Atuação em Gestão de Saúde

Gestão em Saúde é o campo de atuação em saúde relacionado à liderança, gestão e administração de sistemas de saúde públicos e privados, clínicas, laboratórios, hospitais e redes hospitalares nos setores primário, secundário e terciário.

4. Questão 1: Atua com gestão em saúde?

Marcar apenas uma oval.

- ◯ Sim *Pular para a pergunta 5*
- ◯ Não

Tempo de Atuação

Selecionar, em anos, o tempo de atuação em gestão em saúde.

5. Questão 2: Há quanto tempo você atua com gestão em saúde?

Marcar apenas uma oval.

- ◯ Menos de 2 anos
- ◯ 2 a 3 anos
- ◯ 3 a 4 anos
- ◯ 4 a 5 anos
- ◯ 5 a 10 anos
- ◯ 10 anos ou mais

Questão 4: Análise do Grau de Concordância das variáveis relacionadas aos pacientes com maior risco de se tornarem alto custo

Abaixo estão descritas estas mesmas variáveis. Para cada variável, expressar seu grau de concordância em relação a associação destas variáveis com o desfecho "alto custo". Caso discorde ou discorde totalmente, um campo de justificativa será disponibilizado na sequência.

7. Concordância Variáveis X Desfecho

Marque todas que se aplicam.

	Concordo totalmente	Concordo	Indiferente	Discordo	Discordo totalmente
histórico pessoal de transtorno bipolar	☐	☐	☐	☐	☐

8. Justificativa (caso tenha discordado ou discordado totalmente) *

Caso tenha concordado, escrever apenas "sem justificativa"

9. Concordância Variáveis X Desfecho

Marque todas que se aplicam.

	Concordo totalmente	Concordo	Indiferente	Discordo	Discordo totalmente
uso contínuo de medicamento	☐	☐	☐	☐	☐

10. Justificativa (caso tenha discordado ou discordado totalmente) *

Caso tenha concordado, escrever apenas "sem justificativa"

11. Concordância Variáveis X Desfecho

Marque todas que se aplicam.

	Concordo totalmente	Concordo	Indiferente	Discordo	Discordo totalmente
quantidade de refeições realizadas diariamente	☐	☐	☐	☐	☐

12. Justificativa (caso tenha discordado ou discordado totalmente) *

Caso tenha concordado, escrever "sem justificativa"

13. Concordância Variáveis X Desfecho *

Marque todas que se aplicam.

	Concordo Totalmente	Concordo	Indiferente	Discordo	Discordo Totalmente
histórico de dores fortes	☐	☐	☐	☐	☐

27. Questão 5: De acordo com a sua vivência profissional, o que é desejável como elemento de apoio à decisão em gestão de saúde? Uma melhor performance do algoritmo, indicando com maior precisão quem serão os pacientes com risco de se tornarem alto custo, sem explicação de como chegou ao seu resultado, ou uma performance ligeiramente inferior, em que é possível identificar quais variáveis estão mais fortemente associadas a este desfecho?

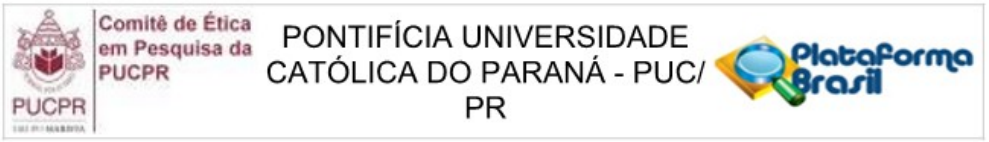
Marcar apenas uma oval.

- ◯ Melhor performance com baixa interpretação
- ◯ Performance inferior com possibilidade de interpretação
- ◯ Nenhuma das alternativas

Agradecemos a sua participação neste estudo.

ANEXO A

PARECER DO COMITÊ DE ÉTICA EM PESQUISA



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Identificação de Pacientes de Alto Custo com a Utilização de Dados Administrativos e Autorreferidos por meio do Aprendizado de Máquina

Pesquisador: Deborah Ribeiro Carvalho

Área Temática:

Versão: 1

CAAE: 46628421.3.0000.0020

Instituição Proponente: Pontifícia Universidade Católica do Parana - PUCPR

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.756.053

Apresentação do Projeto:

Extraído de PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1715679.PDF postado em 10/05/21.

Devido a mudança no perfil epidemiológico das populações, em especial com o aumento da prevalência das doenças crônicas, os sistemas de saúde têm enfrentado desafios para oferecer saúde de forma sustentável, com qualidade e focada no paciente (SALISBURY et al, 2011). Diante disto, é importante o olhar para aqueles pacientes que mais demandam por serviços de saúde, conhecidos como pacientes com alto custo (SMEETS et al, 2020), que compreendem entre 1 a 20% dos pacientes atendidos pelos sistemas de saúde, são responsáveis por consumir mais de 50% dos recursos e precisam de intervenções específicas para atender suas demandas e evitar desperdícios (WAMMES et al, 2019).A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) tem como prioridade as intervenções com foco nestes pacientes, devido ao seu potencial de efetivamente conter o rápido crescimento dos custos com saúde. Portanto, uma abordagem válida, confiável e implementável para prever com precisão quem serão os pacientes com este perfil de alto consumo de recursos é bastante importante para projetar ações sensíveis à redução de custos (TANKE et al, 2019).Muitas vezes estes pacientes recebem um cuidado inapropriado e desnecessário para a severidade de suas doenças, o que comprova a necessidade de conhecimento profundo desta população (SIEKMAN & HILGER, 2018). Identificá-los precocemente por meio de modelos preditivos pode evitar desfechos indesejados e garantir um melhor planejamento

terapêutico e financeiro (MOTURU et al, 2008).Estes modelos dependem majoritariamente de dados administrativos, com base em: diagnósticos, histórico de utilização dos serviços e seus respectivos custos (FERVER et al, 2009). São dados com objetivo de pagamento a prestadores (CARVALHO et al, 2015) que não contemplam dados clínicos ou psicossociais (FERVER et al, 2011). O fato de no Brasil não ser obrigatório o fornecimento da classificação internacional das doenças em guias ambulatoriais, dificulta ainda mais a identificação das características epidemiológicas neste cenário (CARVALHO et al, 2015).Não obstante, é importante considerar fatores comportamentais e socioeconômicos para a identificação de pacientes com alto custo, pois torna os métodos mais efetivos, ampliando a qualidade das predições, dada a possibilidade destes fatores estarem associados a um maior custo (BATES et al, 2014). Dados referentes a estes aspectos não estão presentes nas bases de dados administrativos, porém podem ser coletadas por meio de questionários de autoavaliação de saúde (BOSCARDIN et al, 2015).Estudos realizados a partir de questionários em populações americanas, demonstram que os dados autorreferidos contribuem para a predição de custos (PERRIN et al., 2011; DE SALVO et al., 2009). Com a revolução digital, estes dados passaram a ser coletados com maior facilidade, por meio de websites ou aplicativos em aparelhos celulares, gerando uma grande quantidade de informações de saúde (WERE et al, 2016).Os avanços tecnológicos não se destacam somente na coleta dos dados, mas também nas análises destes, com o advento das técnicas oriundas da inteligência artificial (SUSHIMITA et al, 2015). Dentre elas, estão os algoritmos de machine learning de aprendizado supervisionado (MORID et al, 2017). Neste tipo de aprendizado de máquina, o desfecho de um conjunto de dados é conhecido, existindo um valor da variável resposta a ser predito, ou seja, no conjunto de dados estão disponíveis as variáveis preditoras e a variável de interesse, responsável por guiar a análise (HASTIE et al, 2008). Este tipo de aprendizado se destaca para a predição de pacientes de alto custo (BERTSIMAS et al, 2008) por minimizarem as limitações dos testes estatísticos.Embora os modelos estatísticos, principalmente os modelos de regressão, tenham êxito em suas predições, apresentam alguns desafios importantes: o primeiro é a capacidade limitada de trabalhar com várias variáveis independentes e suas fortes correlações, o que gera multicolinearidade (CECHULIN et al, 2014). O segundo, compreende a natureza dos dados de saúde, em que valores diversos estão presentes, tornando sua distribuição assimétrica. É comum os dados de saúde apresentarem valores extremos, com cauda a direita e, apesar de avanços nas técnicas estatísticas para acomodar esta distribuição, este método não é capaz de performar melhor que o aprendizado supervisionado (MORID et al, 2017).No trabalho conduzido por Chechulin e colaboradores (2014) com o objetivo de predizer pacientes com risco de se

tornarem alto custo no Canadá por meio de uma análise de regressão logística, os autores reiteram os pressupostos acima como limitação metodológica: a grande quantidade de variáveis e os numerosos requisitos relacionados aos dados para execução do modelo, que são facilmente mitigadas pelo emprego de algoritmos de machine learning.Quanto as ações empregadas para o gerenciamento do custo e das condições de saúde de pacientes com alto custo ou, ainda, daqueles que serão alto custo no futuro, é necessário o desenvolvimento de modelos de cuidado com base nas necessidades médicas e socioculturais, bem como suas preferências. Transformar modelos em ganhos clínicos e financeiros exigirá bancos de dados bastante abrangentes com informações de pagamento, clínicas e demais fatores que podem determinar o estado de saúde dos indivíduos (KHULLAR et al, 2018; SMITH et al, 2019).

Hipótese:
Os dados autorreferidos trazem características do estado de saúde dos pacientes que não são identificados e/ou extraídos das bases de dados administrativos. A pesquisa tem como hipótese que o enriquecimento da base de dados com dados autorreferidos pode aumentar a performance preditiva do modelo e trazer informações relevantes do estado de saúde, auxiliando na melhor gestão terapêutica e de recursos destinados a estes pacientes.

Metodologia Proposta:
Trata-se de uma pesquisa quantitativa, retrospectiva e de caráter descritivo. A população da pesquisa é composta por pacientes titulares de planos de saúde coletivos empresariais que responderam um questionário de autoavaliação de saúde. O cenário desta pesquisa é uma empresa contratante de um plano de saúde coletivo empresarial. A pesquisa será realizada em quatro etapas. A etapa 1 compreende o pré-processamento dos dados, em que serão selecionados os dados das bases de dados administrativos e do questionário de saúde a fim de criar uma única base de dados com variáveis de interesse. Os dados administrativos serão utilizados para apurar o custo total de cada paciente, por meio da soma dos custos de todos os procedimentos realizados no período e, também, para quantificar o número de internações, exames, terapias, uso de órtese/prótese e/ou material especial (OPME), e consultas realizadas. Para definição dos pacientes com alto custo será feita uma análise descritiva dos dados para obter o percentual de corte, que deve variar entre 1 e 20% do total de pacientes. Os demais, serão classificados como baixo custo. As variáveis extraídas do questionário de autoavaliação de saúde são: idade, gênero, peso e altura (utilizados para cálculo do índice de massa corpórea), histórico de doenças, estado de saúde

mental, uso de medicamentos, consumo de álcool com avaliação da frequência e quantidade, consumo de tabaco com avaliação da frequência e quantidade, presença de dor e prática de atividades físicas. Em relação aos hábitos, considerados fatores de risco modificáveis para doenças, foram também questionadas as intenções para mudanças e possíveis fatores impeditivos para que estas mudanças ocorram. Considerando a heterogeneidade dos pacientes com alto custo (Blumenthal et al, 2016; Smeets et al, 2020), também serão criadas bases de dados seguindo os mesmos critérios descritos acima, porém, selecionando apenas os pacientes que referem hipertensão arterial sistêmica (HAS) e diabetes mellitus (DM), a fim de avaliar o modelo de identificação de pacientes propensos ao alto custo dentro destas categorias. A etapa 2 será a implementação de algoritmos de machine learning no Waikato Environment for Knowledge Analysis (WEKA) e em linguagem phyton. Na etapa 3, composta pela análise dos dados pós-processados, serão avaliados os resultados considerando as métricas comumente utilizadas para algoritmos de classificação. Nesta etapa, também, serão feitas comparações dos resultados obtidos com o agrupamento de pacientes e aqueles hipertensos e diabéticos, com o objetivo de avaliar se a separação por patologia contribui para um melhor desempenho do modelo. Na etapa 4, será realizada a avaliação por especialistas para identificar o grau de concordância destes profissionais em relação aos resultados obtidos. A busca dos especialistas será realizada por meio da metodologia bola de neve, em que os especialistas com características desejáveis para o estudo recrutam futuros especialistas, informando ao pesquisador seu nomes e endereços eletrônicos. Considerando o atual cenário pandêmico, a comunicação e envio da avaliação serão feitos eletronicamente. Uma vez declarado o aceite no TCLE, o especialista passará a responder as perguntas que compreendem: avaliar a atuação com gestão em saúde, tempo de atuação, formação acadêmica, uma análise do grau de concordância em relação as variáveis associadas ao desfecho de alto custo e uma avaliação sobre importante para a tomada de decisão sobre informações para tomada de decisão em gestão em saúde. O modelo da avaliação pode ser acessado pelo link https://docs.google.com/forms/d/e/1FAIpQLSftzqG05RNx5FihnHwsZCBCuTY7q16TLjNNoP9-eqCEHnXqwQ/viewform.

Os resultados incluídos neste modelo de avaliação são genéricos e simulados, apenas para compreensão de como estarão disponíveis para avaliação dos especialistas.

Critério de Inclusão: Como critérios de inclusão do especialista participante da pesquisa, o profissional deve possuir no mínimo dois anos de experiência em gestão de saúde, não sendo necessário coordenar ou gerenciar equipes, com cargo gerencial e, sim, ter experiência em gestão administrativa em saúde. Para garantir um olhar multidisciplinar a seleção envolverá a busca por

médicos e enfermeiros. A recomendação de especialistas sofre variações na literatura e, para esta pesquisa, será seguida a recomendação proposta por Lynn (1986), que define um número mínimo de cinco e máximo de dez. A busca dos especialistas será realizada através de amostragem não probabilística, denominada bola de neve. Esse método é aplicado quando o acesso a especialistas com características específicas é considerado restrito. No método bola de neve, os especialistas com características desejáveis para o estudo recrutam futuros especialistas entre sua rede de contatos (NADERIFAR; GOLI; GHALJAIE, 2017; SEDGWICK, 2013).

Critério de Exclusão: Não serão selecionados profissionais que atuam na instituição onde os dados foram coletados. Por se tratar de metodologia bola de neve, caso algum especialista da instituição responda o questionário de avaliação, suas respostas serão desconsideradas.

Objetivo da Pesquisa:

Extraído de PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1715679.PDF postado em 10/05/21.

Objetivo Primário: Identificar pacientes de alto custo a partir de dados administrativos, comportamentais e socioeconômicos por meio do aprendizado de máquinas

Objetivo Secundário: Propor um modelo a partir de algoritmos de aprendizado de máquina; Avaliar o modelo por meio de entrevistas com especialistas.

Avaliação dos Riscos e Benefícios:

Extraído de PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1715679.PDF postado em 10/05/21.

Riscos:

Uma vez que os dados utilizados são anonimizados, não há risco de caracterização dos pacientes ou de exposição de dados sensíveis. A proposta de utilização dos resultados para gerenciar a respectiva condição saúde, propor acompanhamento e inserção em linhas de cuidados irão mitigar os riscos dos resultados onerarem ainda mais os custos daqueles pacientes com potencial alto custo. Quanto ao participante (especialista) há o risco de contato pessoal, que será mitigado pelo contato ser estabelecido por envio do link eletrônico do TCLE e do formulário para a avaliação. Quanto ao risco de eventual constrangimento em responder alguma questão o participante (especialista) poderá não responder à alguma questão ou mesmo interromper a participação,

retirando o TCLE a qualquer momento.

Benefícios:

Os pacientes cujos dados serão utilizados nesta pesquisa poderão se beneficiar de ações e programas de saúde propostos pela empresa onde trabalham, mas não serão beneficiados individualmente devido a anonimização das bases de dados. Os especialistas não terão benefícios diretos, porém obterão informações atualizadas sobre identificação precoce de pacientes com alto custo, além de poderem vislumbrar o potencial do uso da inteligência artificial para a gestão de saúde.

Comentários e Considerações sobre a Pesquisa:

Trata-se de projeto de dissertação de mestrado apresentado ao programa de pós-graduação em Tecnologia em Saúde da Pontifícia Universidade Católica do Paraná da linha de pesquisa em Informática em Saúde.

Considerações sobre os Termos de apresentação obrigatória:

Ver "Conclusões ou Pendências e Lista de Inadequações".

Recomendações:

Sem recomendações.

Conclusões ou Pendências e Lista de Inadequações:

Não foram observados óbices de ordem ética para a execução da proposta conforme apresentada. Projeto de pesquisa aprovado, pois em consonância com os ditames éticos e legais das Resoluções nºs 466/12 e 510/16, ambas do CNS.

Considerações Finais a critério do CEP:

Lembramos aos senhores pesquisadores que, no cumprimento da Resolução 466/12, o Comitê de Ética em Pesquisa (CEP) deverá receber relatórios anuais sobre o andamento do estudo, bem como a qualquer tempo e a critério do pesquisador nos casos de relevância, além do envio dos relatos de eventos adversos, para conhecimento deste Comitê.

Salientamos ainda, a necessidade de relatório completo ao final do estudo. Eventuais modificações ou emendas ao protocolo devem ser apresentadas ao CEP-PUCPR de forma clara e sucinta,

identificando a parte do protocolo a ser modificado e as suas justificativas.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1715679.pdf	10/05/2021 14:33:53		Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_PPGTS.docx	10/05/2021 14:33:31	ANA LUISA GONCALVES GOMES COELHO SELEME	Aceito
Projeto Detalhado / Brochura Investigador	brochura_CEP.docx	06/05/2021 18:53:53	ANA LUISA GONCALVES GOMES COELHO	Aceito
Outros	Formulario_Especialistas.pdf	06/05/2021 18:16:15	ANA LUISA GONCALVES GOMES COELHO	Aceito
Folha de Rosto	folhaderosto_assinada.pdf	29/04/2021 09:42:33	ANA LUISA GONCALVES GOMES COELHO	Aceito
Cronograma	CRONOGRAMA.docx	22/04/2021 09:28:30	ANA LUISA GONCALVES GOMES COELHO	Aceito
Orçamento	ORCAMENTO.docx	22/04/2021 09:27:34	ANA LUISA GONCALVES GOMES COELHO	Aceito
Declaração de Pesquisadores	TCUD_assinado.pdf	12/04/2021 18:19:37	ANA LUISA GONCALVES GOMES COELHO	Aceito
Declaração de Instituição e Infraestrutura	carta_dados.pdf	12/04/2021 12:09:08	ANA LUISA GONCALVES GOMES COELHO	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

Continuação do Parecer: 4.756.053

CURITIBA, 06 de Junho de 2021

Assinado por:
Ana Carla Efing
(Coordenador(a))

Endereço: Rua Imaculada Conceição 1155
Bairro: Prado Velho **CEP:** 80.215-901
UF: PR **Município:** CURITIBA
Telefone: (41)3271-2103 **Fax:** (41)3271-2103 **E-mail:** nep@pucpr.br

ANEXO B

QUESTIONÁRIO DE AVALIAÇÃO GLOBAL DE SAÚDE

Nome Completo: _____
Empresa: _____ Matrícula: _____
Data de Nascimento: ____/____/_____ Sexo: ( ) Masculino ( ) Feminino CPF _____
Nome da Unidade: _____
Nome do Setor: _____ Nome do Cargo: _____
Telefone para contato: _____

1. Autorizo a equipe de saúde a ter acesso às suas respostas para ações em saúde? ( ) Sim ( ) Não
2. Quais dos benefícios que a empresa oferece para você gratuitamente você conhece? ( ) Psicoterapia Online gratuita (4 sessões por mês) ( ) Implus App com orientações em saúde 24 horas ( ) Implus App com programa de desconto em medicamentos ( ) Implus App com orientações de nutricionista e educador físico ( ) Implus Monitor para checagem de sintomas de COVID-19 e monitoramento de sintomáticos

3. Você classificaria seu estado de saúde como: ( ) muito bom ( ) bom ( ) regular ( ) ruim ( ) muito ruim
4. Quão satisfeito você está com sua qualidade de vida? ( ) muito satisfeito ( ) Satisfeito ( ) Neutro ( ) Insatisfeito ( ) Muito insatisfeito
5. Qual meio de informação preferível para se informar sobre a saúde? ( ) Aulas ( ) Email ( ) WhatsApp, mensagem de texto ( ) Impressos (livros/jornais/folhetos) ( ) Aplicativos de celular
6. Estado civil: ( ) solteiro ( ) casado ( ) separado ( ) outra situação
7. Escolaridade: ( ) Ensino Fundamental ( ) Ensino Médio ( ) Ensino superior (incompleto) ( ) Ensino superior (completo) ( ) Pós-graduação

HISTÓRIO MÉDICO PESSOAL

8. Meu peso atual: _____ Minha altura:_____
9. Faz consultas médicas regularmente (1 vez ao ano, pelo menos)? ( ) Sim ( ) Não
10. Suas vacinas estão atualizadas? ( ) Não ( ) Não sei ( ) Sim
11. Faz uso de medicamento todos os dias? ( ) Sim ( ) Não. Se sim, qual(is)? _____
12. Tem ou teve o diagnóstico de:
    - Hipertensão arterial? ( ) Não ( ) Sim, mas não faço uso de medicamento ( ) Sim, e uso medicamento regularmente
    - Diabetes? ( ) Não ( ) Sim, mas não faço uso de medicamento ( ) Sim, e faço tratamento apenas com comprimidos ( ) Sim, e faço tratamento com insulina
    - Hipercolesterolemia (colesterol alto)? ( ) Não ( ) Sim, mas não faço uso de medicamento ( ) Sim, e uso medicamento regularmente
13. Assinale qual(is) diagnósticos você apresenta/apresentou:

( ) Alergias de pele/medicamentos	( ) Epilepsia
( ) Anemia falciforme	( ) Hepatite B ou C, HIV
( ) Ansiedade	( ) Artrose/Artrite
( ) Depressão	( ) Dor crônica
( ) Transtorno Afetivo Bipolar	( ) Doenças de estômago (gastrite, úlcera, refluxo)
( ) Asma/bronquite	( ) Doenças de intestino
( ) Doença Pulmonar Obstrutiva Crônica (DPOC) ou Enfisema	( ) Doenças Reumáticas
( ) Doenças de rim (doença renal crônica)	( ) Insônia
( ) Derrame (AVC)	( ) Rinite
( ) Doença de tireoide (hipo ou hipertireoidismo)	( ) Problemas de audição
( ) Infarto, angina ou arritmia	( ) Problemas de visão (astigmatismo, miopia, hipermetropia, glaucoma, entre outros)
( ) Insuficiência Cardíaca Congestiva (ICC)	( ) Osteoporose
( ) Dor de cabeça frequente	( ) Sinusite crônica
( ) Obesidade (IMC ≥ 30)	( ) Câncer
( ) Transplante de órgão (em estado de imunossupressão)	( ) Não tenho doenças
( ) Doenças do sangue	

Outras: _____

HISTÓRICO MÉDICO FAMILIAR

14. Assinale de algum parente de primeiro grau (pai, mãe, irmãos biológicos ou filhos) já recebeu o diagnóstico de:
    - ( ) Diabetes ( ) Hipertensão arterial ( ) Colesterol e/ou triglicerídeos alto ( ) Derrame (AVC) ou infarto precoce (Pai ou irmão biológico, com menos de 55 anos; Mãe ou irmã biológica, antes de 65 anos)
    - Algum dos cânceres a seguir: ( ) Câncer de mama ( ) Câncer de Intestino (Cólon) ( ) Câncer de Próstata ( ) Câncer de Pulmão ( ) Câncer do colo do útero ( ) Câncer de pele
    - ( ) Nenhuma das alternativas anteriores

SAÚDE MENTAL

15. Como você considera seu estado de saúde mental hoje? ( ) Ótimo ( ) Bom ( ) Regular ( ) Ruim
16. Assinale quais frases refletem como você tem se sentido nos últimos 30 dias:
    ( ) Tenho dores de cabeça frequentes ( ) Tenho falta de apetite ( ) Durmo mal ( ) Assusto-me com facilidade ( ) Tenho tremores de mãos ( ) Sinto-me nervoso, tenso ou preocupado ( ) Tenho má digestão ( ) Tenho dificuldades para pensar com clareza ( ) Tenho me sentido triste ultimamente ( ) Tenho chorado mais que o costume ( ) Encontro dificuldades para realizar com satisfação minhas atividades diárias ( ) Tenho dificuldades em tomar decisões ( ) Tenho dificuldades no serviço (meu trabalho é penoso, me causa sofrimento) ( ) Sou incapaz de desempenhar um papel útil na minha vida ( ) Tenho perdido interesse pelas coisas ( ) Sinto-me inútil, sem préstimo ( ) Tenho tido ideias de acabar com minha vida ( ) Sinto-me cansado o tempo todo ( ) Tenho sensações desagradáveis no estômago ( ) Canso-me com facilidade
    ( ) Nenhuma das anteriores
17. Nos últimos 6 meses, tem atravessado um período significativamente estressante? ( ) Sim ( ) Não
18. Na sua história de vida, você enfrentou algum evento potencialmente ameaçador à sua vida tais como catástrofes naturais, acidentes graves, violência física ou sexual, combate militar ou abuso infantil? ( ) Sim ( ) Não
19. Faz acompanhamento com psiquiatra? ( ) Sim ( ) Não
20. Faz acompanhamento com psicólogo? ( ) Sim ( ) Não
21. Faz uso de medicamento controlado? ( ) Sim ( ) Não

SONO

22. Em relação ao sono, assinale:
    Quantas horas costuma dormir por noite? ( ) 4 horas ou menos ( ) 5 ou 6 horas ( ) 7 horas ou mais
    Sente-se descansado após as horas de sono? ( ) Sim ( ) Não
    Qual a chance de você cochilar sentado e lendo: ( ) Nenhuma ( ) Pequena ( ) Moderada ( ) Alta
    Qual a chance de você cochilar lugar público: ( ) Nenhuma ( ) Pequena ( ) Moderada ( ) Alta
    Qual a chance de você cochilar sentado após o almoço: ( ) Nenhuma ( ) Pequena ( ) Moderada ( ) Alta
    Qual a chance de você cochilar no trânsito: ( ) Nenhuma ( ) Pequena ( ) Moderada ( ) Alta

HÁBITOS DE VIDA

TABAGISMO

23. É tabagista?
    ( ) Sim. Há quantos anos? _____ Quantos cigarros por dia? _____
    ( ) Não, nunca fumou
    ( ) Não, e parou há menos de 1 ano
    ( ) Não, e parou há mais de 1 ano
24. Qual o grau de motivação para mudar os hábitos em relação ao cigarro?
    ( ) Não se aplica
    ( ) Não penso nisso/Não vejo necessidade de mudança nesse momento
    ( ) Penso um pouco em mudar
    ( ) Penso muito em mudar
    ( ) Estou tomando atitudes para mudar

ETILISMO

25. Faz uso de bebida alcoólica? ( ) Sim ( ) Não, nunca consome bebida alcoólica
26. Nos últimos 30 dias, quantas vezes ingeriu bebida alcoólica? ( ) Nenhuma vez ( ) 1 vez ao mês ou menos ( ) 2 a 4 vezes ao mês ( ) 1 a 3 vezes na semana ( ) 4 ou mais vezes na semana
27. Qual o grau de motivação para mudar os hábitos em relação à bebida alcoólica?
    ( ) Não se aplica

( ) Não penso nisso/Não vejo necessidade de mudança nesse momento
( ) Penso um pouco em mudar
( ) Penso muito em mudar
( ) Estou tomando atitudes para mudar

ATIVIDADE FÍSICA

28. Você tem alguma condição física ou de saúde que o impeça de realizar algum exercício físico ou esporte? ( ) Sim ( ) Não
29. Quantas vezes por semana você realiza 20 minutos ou mais de **atividade física VIGOROSA** que faça você suar ou ofegar? (por exemplo: musculação, correr, carregar objetos pesados, atividade aeróbica, andar de bicicleta rapidamente, esportes em grupo, natação, dançar)
    ( ) 5 ou mais vezes na ( ) 3 a 4 vezes na semana ( ) 1 a 2 vezes na semana ( ) nenhuma vez na semana
30. Quantas vezes por semana você realiza 30 minutos de **atividade física MODERADA ou CAMINHADA** que aumente a frequência cardíaca ou a respiração? (por exemplo: atividades de cuidado com a casa, lavar o carro, andar de bicicleta, brincar com crianças, subir escadas)
    ( ) 5 ou mais vezes na semana ( ) 3 a 4 vezes na semana ( ) 1 a 2 vezes na semana ( ) nenhuma vez na semana
31. Qual o grau de motivação para mudar os hábitos em relação à atividade física?
    ( ) Não se aplica
    ( ) Não penso nisso/Não vejo necessidade de mudança nesse momento
    ( ) Penso um pouco em mudar
    ( ) Penso muito em mudar
    ( ) Estou tomando atitudes para mudar

ALIMENTAÇÃO

32. Assinale quais refeições realiza diariamente: ( ) café da manhã ( ) lanche da manhã ( ) almoço ( ) lanche da tarde ( ) jantar ( ) ceia
33. Qual o seu consumo de água por dia? ( ) Até 500ml ( ) 500 a 1000ml ( ) 1000 a 1500ml ( ) Mais de 1500ml
34. Sobre alimentação, com que frequência consome:
    - Arroz, pão, macarrão, cereais, bolachas: ( ) Nunca ( ) Raramente ( ) 1 a 3 dias na semana ( ) 4 a 7 dias na semana
    - Frutas: ( ) Nunca ( ) Raramente ( ) 1 a 3 dias na semana ( ) 4 a 7 dias na semana
    - Verduras e legumes: ( ) Nunca ( ) Raramente ( ) 1 a 3 dias na semana ( ) 4 a 7 dias na semana
    - Carnes, feijão e ovos: ( ) Nunca ( ) Raramente ( ) 1 a 3 dias na semana ( ) 4 a 7 dias na semana
    - Leite, queijos, requeijão e iogurtes: ( ) Nunca ( ) Raramente ( ) 1 a 3 dias na semana ( ) 4 a 7 dias na semana
    - Manteiga, margarina, frituras, embutidos, azeite, carnes gordurosas: ( ) Nunca ( ) Raramente ( ) 1 a 3 dias na semana ( ) 4 a 7 dias na semana
    - Achocolatados, refrigerantes, bolos, chocolates, comidas doces, sucos artificiais (de pacotinho): ( ) Nunca ( ) Raramente ( ) 1 a 3 dias na semana ( ) 4 a 7 dias na semana
35. Qual o grau de motivação para mudar os hábitos em relação à alimentação?
    ( ) Não se aplica
    ( ) Não penso nisso/Não vejo necessidade de mudança nesse momento
    ( ) Penso um pouco em mudar
    ( ) Penso muito em mudar
    ( ) Estou tomando atitudes para mudar

SAÚDE PREVENTIVA

36. Maiores de 50 anos: Já realizou o exame de sangue oculto nas fezes? ( ) Sim ( ) Não/Não se aplica (menores de 50 anos)
37. Mulheres 25 -64 anos: exame preventivo do câncer do colo do útero no último ano? ( ) Sim ( ) Não/Não se aplica (homens)
38. Mulheres de 50-69: mamografia para prevenção do câncer de mama? ( ) Sim, há menos de 1 ano ( ) Sim, entre 1 e 2 anos ( ) Sim, há mais de 2 anos ( ) Não/Não se aplica (homens)
39. Mulheres: está gestante? ( ) Sim ( ) Não/Não se aplica (homens)
40. Homens mais de 50 anos: realizou consulta médica para avaliação da sua saúde?
    ( ) menos de 1 anos ( ) entre 1 e 2 anos ( ) mais de 2 anos ( ) não/não se aplica (mulheres)
41. Homens e mulheres: Qual o método utilizado para Planejamento Familiar?
    ( ) Vasectomia ( ) Camisinha masculina ( ) Camisinha feminina ( ) Laqueadura ( ) DIU ( ) Anticoncepcional oral ( ) Anticoncepcional injetável ( ) Outros métodos

OSTEOMUSCULAR

42. Já realizou alguma cirurgia ortopédica? ( ) Não ( ) Sim, na coluna/costas ( ) Sim, no ombro/braço ( ) Sim, na perna/joelho/tornozelo ( ) Sim, na mão/punho/dedos
43. Assinale onde sente dor forte: ( ) Não sinto dor forte ( ) Cabeça ( ) Pescoço ( ) Ombros ( ) Braços ( ) Mão, punho e dedos ( ) Costas/lombar ( ) Quadril ( ) Perna ( ) Joelho ( ) Tornozelo e pés

SAÚDE BUCAL

44. Quanto a Saúde bucal assinale:
    - Quantas vezes ao dia você costuma escovar os dentes: ( ) 1 a 2 vezes ( ) 3 a 4 vezes ( ) 5 vezes ou mais
    - Realizou consulta com dentista nos últimos 12 meses: ( ) Sim ( ) Não
    - Apresenta sangramento de gengiva ao escovar os dentes: ( ) Sim ( ) Não
    - Dor de dente: ( ) Sim ( ) Não
    - Manchas brancas ou vermelhas na boca: ( ) Sim ( ) Não

TRABALHO REMOTO

45. Você está em trabalho remoto? ( ) Sim ( ) Não
46. Se sim, como você avalia essa modalidade de trabalho? ( ) Excelente ( ) Bom ( ) Regular
47. Como você avalia sua situação ergonômica nas atividades em trabalho remoto? ( ) Excelente ( ) Boa ( ) Regular