

UNIVERSIDADE FEDERAL DE MINAS GERAIS
BACHARELADO EM CIÊNCIAS ATUARIAIS

Pedro Vitor Bernardes Brandão

**MODELAGEM PREDITIVA DE INSUFICIÊNCIA RENAL CRÔNICA:
COMPARAÇÃO DAS TÉCNICAS DE REGRESSÃO LOGÍSTICA E ÁRVORE DE
CLASSIFICAÇÃO**

Belo Horizonte

2022

PEDRO VITOR BERNARDES BRANDÃO

**MODELAGEM PREDITIVA DE INSUFICIÊNCIA RENAL CRÔNICA:
COMPARAÇÃO DAS TÉCNICAS DE REGRESSÃO LOGÍSTICA E ÁRVORE DE
CLASSIFICAÇÃO**

Trabalho de conclusão de curso apresentado ao Curso Bacharelado em Ciências Atuariais da Universidade Federal de Minas Gerais para obtenção do grau de bacharel em Ciências Atuariais.

Orientadora: Ilka Afonso Reis.

Belo Horizonte

2022

AGRADECIMENTOS

Agradeço a Deus que depositou em mim a fé, a esperança e a certeza de que meus sonhos se realizariam e que meus esforços não seriam em vão.

Agradeço a meus pais Vânia e Paulo, irmão Paulo e a Lais, por estarem comigo nesta caminhada intensa de estudos, sempre me apoiando e me incentivando, para que fosse mais longe e alcançasse resultados cada vez maiores.

Agradeço aos professores do curso que se dedicaram em passar seu conhecimento e nos incentivar enquanto estudantes a sempre galgar lugares altos e de prestígio.

Um agradecimento aos atuários que tive o prazer de trabalhar junto e me ensinaram o “ser atuário”, os desafios da profissão, foram companheiros na jornada de faculdade e estágio, me apoiando e incentivando sempre. Em especial ao Rafael Esteves e Tatiana Xavier, que estiveram presentes na escrita e idealização do trabalho aqui apresentado, não medindo esforços em tirar parte do seu tempo de descanso para me auxiliar e incentivar.

Deixo um agradecimento especial a Ilka, minha orientadora, pelo incentivo, dedicação e por todo apoio e paciência ao longo da elaboração de todo trabalho.

RESUMO

Tendo em vista o aumento dos gastos assistenciais nas carteiras de planos de saúde, o mercado converge para medidas de promoção e prevenção de riscos e doenças. Este trabalho propõe a utilização de dois modelos: regressão logística e árvore de classificação para predição de Insuficiência Renal Crônica, uma doença que possui alto custo de tratamento. Para realização do estudo, foram utilizadas as informações hospitalares e ambulatoriais do Sistema Único de Saúde (SUS). A fim de avaliar o poder preditivo dos modelos, foram utilizadas medidas de sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e a área sob a curva ROC. A partir da aplicação das técnicas, foi possível identificar uma ligeira superioridade da regressão logística em relação a árvore de classificação. Além disso, foi possível evidenciar como o desbalanceamento dos dados prejudica a predição de insuficiência renal crônica.

Palavras-chave: Saúde Suplementar, Sistema Único de Saúde, Machine Learning, Regressão Logística.

ABSTRACT

In view of the increase in healthcare expenses in the health plan portfolios, the market converges towards measures to promote and prevent risks and diseases. This project uses two models of efficient work, a classification tree of Chronic Kidney Failure that has a high cost of treatment. To carry out the study, hospital and outpatient information from the Unified Health System (SUS) was used. In order to assess the predictive power of the models, measurements of sensitivity, specificity and an area under the ROC curvature were used, using the Kappa coefficient. From the application of the techniques, a slight superiority was possible in addition to allowing the reproduction as a function of the classification of this data, it was evidenced as the prediction of chronic renal function and as the settings for adjustments to improve the data consubstantial in the predictive power of the models.

Keywords: Supplementary Health, Unified Health System, Machine Learning, Logistic Regression.

LISTA DE TABELAS

Tabela 1 – Despesas por ano e grupo de procedimentos (2018, 2019 e 2020).....	22
Tabela 2 – Categorias de CID com as 10 maiores quantidades de utilizações do SUS, de 2008 a 2018	52
Tabela 3 – Categorias de CID com os 10 maiores custos de tratamento do SUS, de 2008 a 2018	52
Tabela 4 – Proporção de usuários do SUS diagnosticados com IRC por estado, de 2008 a 2018	54
Tabela 5 – Procedimentos realizados por usuários diagnosticados com IRC, de 2008 a 2018	55
Tabela 6 – Diagnóstico por Capítulos do CID dos usuários do SUS antes de adoecerem com IRC, de 2008 a 2018.....	56
Tabela 7 – Medidas de qualidade dos modelos de regressão logística e de árvore de decisão sob os diferentes pontos de corte	58

LISTA DE FIGURAS

Figura 1 – Estrutura da árvore de classificação	30
Figura 2 – Base hipotética para análise	38
Figura 3 – Limpeza da base de análise.....	39
Figura 4 – Criação da variável Indicador na base de análise	40
Figura 5 – Aspecto geral da base final	42

LISTA DE GRÁFICOS

Gráfico 1 – Sinistralidade das Operadoras Médico Hospitalares (2018 a 3º Tri 2021)	21
Gráfico 2 – Frequência de utilização em Consultas por beneficiário (2018 e 2019)	23
Gráfico 3 – Frequência de utilização em Exames por beneficiário (2018 e 2019)	23
Gráfico 4 – Frequência de utilização em Internação por beneficiário (2018 e 2019)	24
Gráfico 5 – Função Sigmóide	29
Gráfico 6 – Proporção de utilização do SUS por sexo, Brasil, total de 2008 a 2018	46
Gráfico 7 – Proporção de utilização do SUS por sexo, Brasil, 2008 a 2018	47
Gráfico 8 – Proporção de utilização por Raça/Cor, Brasil, 2008 a 2018	47
Gráfico 9 – Proporção de utilização dos SUS por Nacionalidade, Brasil, 2008 a 2018	48
Gráfico 10 – Distribuição etária dos usuários do SUS no Brasil de 2008 a 2018	49
Gráfico 11 – Complexidade dos procedimentos ambulatoriais, Brasil, 2008 a 2018	50
Gráfico 12 – Caráter da Internação, Brasil, 2008 a 2018	50
Gráfico 13 – Distribuição de frequências dos usuários do SUS diagnosticados com IRC de 2008 a 2018	55
Gráfico 14 – Curva ROC dos modelos de regressão logística e de árvore de classificação	59

LISTA DE QUADROS

Quadro 1 – Descrição das variáveis seleccionas.....	37
Quadro 2 – Grupo de variáveis criadas.....	41
Quadro 3 – Grupo de procedimentos	42
Quadro 4 – Critério de classificação da curva ROC (Hosmer et al., 2013)	45

SUMÁRIO

AGRADECIMENTOS	3
RESUMO.....	4
ABSTRACT	5
LISTA DE TABELAS	6
LISTA DE FIGURAS	7
LISTA DE GRÁFICOS.....	8
LISTA DE QUADROS	9
SUMÁRIO.....	10
1. INTRODUÇÃO.....	12
2. REVISÃO DA LITERATURA.....	14
2.1. Cuidados com saúde no Brasil	14
2.1.1. <i>O modelo de cuidado público no Brasil</i>	16
2.1.2. <i>O modelo de cuidado privado no Brasil</i>	17
2.1.3. <i>Interfaces da promoção e prevenção a saúde nas esferas públicas e privadas</i>	19
2.2. Receitas e despesas, gestões de cuidado eficientes na contenção dos custos 20	
2.2.1. <i>Despesas a serem honradas pelas operadoras de planos de saúde</i>	22
2.3. <i>Outliers na saúde suplementar: modelos para enfrentar os altos gastos dos beneficiários de alto custo</i>	24
2.3.1. <i>Beneficiários de alto custo nas carteiras dos planos de saúde</i>	25
2.3.2. <i>Adoção de modelos preditivos no mercado de saúde suplementar</i>	27
3. MODELOS UTILIZADOS PARA A CLASSIFICAÇÃO	28
3.1. Regressão Logística.....	28
3.2. Árvore de Classificação.....	30
4. METODOLOGIA	34
4.1. Base de dados.....	34
4.2. Análise descritiva dos dados brutos	35
4.3. Pré-processamento das bases de dados	36
4.4. Identificação e histórico do usuário	37
4.5. Montagem da base final	40
4.6. Ambiente Computacional	43
4.7. Avaliação da qualidade dos modelos	44

5. RESULTADOS E DISCUSSÃO	45
5.1. População estudada.....	46
5.2. Perfil do atendimento prestado.....	49
5.3. Perfil dos procedimentos e diagnósticos dos usuários	51
5.4. Escolha do desfecho a ser estudado.....	53
5.5. Perfil da população com Insuficiência Renal Crônica (IRC)	54
5.6. Medidas de qualidade dos modelos	57
6. CONCLUSÃO	60
7. REFERENCIA BIBLIOGRÁFICA	62
APÊNDICE A.....	67
APÊNDICE B.....	72
APÊNDICE C	73

1. INTRODUÇÃO

Ao longo dos anos os brasileiros vêm se preocupando cada vez mais com seu estado de saúde. Uma pesquisa de monitoramento da Hibou (2021), Instituto de Pesquisa e Monitoramento de Mercado, feita com cidadãos com idade superior a 20 anos, de todas as classes sociais, nas 5 regiões do país, mostrou que 90% da população brasileira gostaria de fazer mais pela saúde física e mental em 2021.

No Brasil, o estudo sobre a prestação de assistência médica hospitalar se pauta em dois grandes pilares, o público e o privado. O primeiro se refere as políticas públicas de saúde oferecidas pelo Governo Federal, uma vez que a constituição de 1988 defini a saúde como um dever do estado. No Brasil, todas as políticas públicas de saúde foram incluídas no Sistema Único de Saúde (SUS), que presta atendimento ambulatorial e hospitalar para toda a população brasileira. Já o privado, refere-se ao sistema de atendimento suplementar, ou seja, mediante o pagamento de contraprestações pecuniárias, os indivíduos possuem acesso aos serviços de atendimento médico.

Após anos de tramitação, o marco regulatório da saúde suplementar no Brasil foi finalmente aprovado em 1998, com a publicação da lei 9.656/98, editada sobre forma de Medida Provisória e que atualmente vigora sob a sua 44ª edição, e posteriormente a lei 9.961/00 que criou a ANS. Dentre os pontos que a lei instaurou no mercado de saúde suplementar, destacam-se a) proibição da rescisão por uma das partes dos contratos; b) controle por parte do governo dos reajustes de preços dos planos de saúde individual; c) proibição de seleção de risco por doença ou lesão pré-existente; d) regulamentação das coberturas mínimas obrigatórias (Rol Mínimo de Procedimentos); e) controle atuarial dos preços de venda; f) regras de entrada, operação e saída de operadoras. g) preços limitados pela regra de faixas etárias; e h) regulamentação dos períodos de carência (LEAL, 2007).

Como reflexo dos maiores cuidados dos indivíduos com a saúde, os planos privados médico-hospitalares tiveram um crescimento do número de beneficiários. No ano de 2020 foi observado uma adesão de mais de 1 milhão de pessoas a algum plano de saúde, chegando atualmente à marca de 48 milhões de beneficiários, cerca de 22% da população brasileira, segundo dados do Agência Nacional de Saúde Suplementar (ANS). Diante da inserção em um plano de saúde, o indivíduo é capaz

de estabilizar a variação de sua renda diante de eventos inesperados que possam afetar a sua saúde física e, conseqüentemente, sua saúde financeira (LIMA, 2011). Desta forma, este instrumento de proteção vem sendo necessário, uma vez que os custos de saúde tendem a aumentar com o tempo e os serviços oferecidos pelo setor público ainda se encontram muito abaixo das necessidades da sociedade, o que pauta a discussão da efetiva ação do setor de planos privados, que tem caráter suplementar, porém atua de forma complementar ao SUS.

A base da sustentação do mercado de planos de saúde se faz através do mecanismo da cooperação mútua e voluntária, conhecida como mutualismo, no qual os indivíduos contribuem para que o plano de saúde faça a realocação das receitas no pagamento das despesas médicas de seus beneficiários, podendo ocorrer de algum beneficiário realizar o pagamento de suas contraprestações e não ter nenhuma despesa médica hospitalar, ou seja, ele estará financiando o pagamento de consultas, exames, cirurgias de outra pessoa que está dentro do grupo na qual ele faz parte (LIMA, 2011). Sendo assim, o gerenciamento dos gastos dos beneficiários e o devido cálculo das contraprestações que irão suprir as despesas assistenciais futuras de todos os beneficiários é de suma importância, uma vez que a legislação vigente não permite a seleção dos beneficiários que melhor se enquadrem ao plano de saúde.

Num cenário ótimo, uma operadora de planos de saúde deseja ter beneficiários que utilizem de forma adequada ou baixa os recursos do seu plano, para que seus gastos com as despesas médicas não sejam elevados e influenciem negativamente em sua solvência. Este cenário ótimo ocorre com uma frequência muito baixa, uma vez que os estados de saúde e as necessidades de atendimento médico dos beneficiários são diversos, não tendo a operadora o poder de controlar essas características nos indivíduos que se inserem em sua carteira de beneficiários.

A baixa expansão da oferta de planos de saúde, em consonância à enorme demanda pelos produtos, gerou uma pressão sobre os preços dificultando a entrada de novos consumidores neste setor e expulsando aqueles que tinham um bom nível de saúde e com menor risco assistencial (LEAL, 2007). Isto se dá pelo fato de que as operadoras são obrigadas a tratar as doenças de seus beneficiários de alto custo, estes tratamentos médicos acabam provocando um aumento nas mensalidades dos planos e expulsando os beneficiários que tem gastos baixos e que os preços já não se ajustam as suas restrições orçamentárias.

Segundo Pelletier (1996), o modelo de promoção da saúde é descoberto pelas operadoras nos cenários em que sua carteira começa a ter problemas com alto custos de seus beneficiários. Este fenômeno ocorre devido à assimetria na informação entre a empresa e o consumidor, no que diz respeito ao risco que esse representa para a operadora contratada (LEAL, 2007). A consequência desta assimetria é uma precificação frágil, que será limitado à cobrança de um preço médio para todos os agentes independente dos seus níveis de risco. Tendo em vista esta situação, modelos de saúde preditiva vêm ganhando destaque nas operadoras de assistência médica brasileiras.

Contudo o ponto chave para implementar programas de prevenção a saúde está em identificar quais são as doenças atrelada aos beneficiários que possuem gastos assistências acima da média da carteira, a fim de tratá-los de forma antecipada, antes dos agravamentos das doenças. Esta ação pode ocasionar impactos significativos na redução de custos e um equilíbrio econômico a ser vivenciado pelas operadoras. Os modelos preditivos para identificação desses beneficiários a longo prazo são baseados em dados históricos, dos quais é possível extrair informações que mostrem o real estado da natureza do problema, ou seja, os dados extraídos do meio em que o plano de saúde está exposto mostra a natureza do problema em questão (LOPES, 2018).

Tendo em vista este contexto, o objetivo deste estudo é comparar as técnicas de regressão logística e árvore de classificação para prever a ocorrência de uma doença que possui tratamento mais oneroso para os planos de saúde.

A relevância deste estudo está em identificar quem são os usuários que possuem ou terão gastos elevados, e poder oferecer a eles o tratamento dessas doenças no presente de forma preventiva, a fim de diminuir os custos assistenciais e manter um número cada vez maior de usuários saudáveis dentro da carteira.

2. REVISÃO DA LITERATURA

2.1. Cuidados com saúde no Brasil

A saúde é essencial na vida de qualquer ser humano, desta forma, discutir saúde e suas interfaces se torna fundamental em qualquer esfera e organização social. As demandas por serviços de saúde resultam da conjugação de fatores sociais,

individuais e culturais prevalentes na população (SAWYER et. al., 2002). Falar em saúde, é entender qual é o meio, os hábitos dos indivíduos, a forma de organização e sobretudo o nível de conhecimento e interesse em tratar e encontrar soluções para as diversas problemáticas que hoje apresenta o campo público e privado.

No Brasil, o enfrentamento das problemáticas relacionadas a saúde e os cuidados dos indivíduos surgem antes de que órgãos reguladores ou normativos institucionais de estruturação do setor público e privado de assistência média, venham a existir. Desde a instalação da colônia até a década de 1930, as ações eram desenvolvidas sem nenhum tipo de organização institucional. Segundo Viotti (2017) o recurso da maior parcela da população quando acometida de doenças, em sua maioria se pendia para os profissionais não licenciados, como os curandeiros e parteiras.

Ainda, a autora explica que a preferência da população se resume em três fatores: acessibilidade, custo e efetividade. Na década de 1930, além do tratamento com os médicos terem um custo muito elevado, que parte considerável da população não tinha condições de arcar, os profissionais disponíveis no mercado eram muito poucos. Além disso, os tratamentos eram baseados em um sistema dogmático, que nem sempre apresentava resultados satisfatórios.

Segundo Braga (Braga e Paula, 1985:41-42) a Saúde emerge como “questão social” no Brasil no início do século XX, momento em que foi iniciado, por volta da década de trinta, diversas discussões no nascente movimento operário sobre a saúde como uma reivindicação trabalhista. Os movimentos sindicais estruturam a luta por melhores condições de saúde, acessibilidade aos tratamentos, além de apresentar de forma efetiva uma demanda que o mercado ainda não observava (MENDES et al., 2012).

Tais movimentos provocaram transformações na estrutura de oferta da saúde para os brasileiros, sendo criadas e extintos diversos órgãos de prevenção e controle de doenças, tanto nas esferas públicas e privadas, a fim de proporcionar melhores controles sanitários e atendimento adequado à população.

2.1.1. O modelo de cuidado público no Brasil

No caso brasileiro, nos anos de 1970, concomitantemente ao acelerado crescimento do número de trabalhadores industriais, houve um forte incremento na organização dos trabalhadores em torno da regulamentação da jornada de trabalho e em busca de melhores salários. São também dessa década os primeiros movimentos em defesa da saúde pela melhoria das condições de trabalho (GOMEZ et al., 2018).

Antes de 1988, o sistema público de saúde no Brasil era voltado apenas para os contribuintes da Previdência social. Os cuidados com a saúde eram centralizados e a sua responsabilidade era de órgãos federais, em que os usuários não tinham efetiva participação ou poder de opinião sobre a qualidade e disponibilidade dos serviços prestados (VIOTTI, 2017).

Antes da implementação do SUS, o foco do tratamento era na doença. Neste período, cerca de 30 milhões de pessoas tinham acesso aos sistemas de saúde e tratamento, segundo fonte do Ministério da Saúde. Já as pessoas que não contribuam para a previdência social, muitas vezes, eram marginalizadas, dependiam da caridade ou de filantropias (TEIXEIRA, 2011).

O Sistema Único de Saúde (SUS), surge através do advento da constituição cidadã, em que estabelece a saúde como um “Direito de Cidadania e um dever do Estado”. Nesse sentido, o SUS vem para assumir um órgão com princípios de universalidade, equidade e integralidade da atenção à saúde da população brasileira (TEIXEIRA, 2011). Logo, os pilares do sistema de saúde inaugurado, era atender as demandas de toda a população quanto aos cuidados com a saúde de forma equitativa e integral.

O contraponto a toda ideologia implantada pelo SUS era que a população brasileira carregava em sua sociedade valores que tendem mais para a diferenciação, o individualismo e a distinção do que para a solidariedade, a coletividade e a igualdade (PAIM, 2018). O Sistema então começa a passar por pressões de diversas frentes, seja da população que exige melhores condições de tratamento, de médicos que tem seus interesses destoantes daqueles idealizados pelo SUS, pressões midiáticas e jornalísticas, dentre outras.

A intensa discussão sobre a eficiência do modelo de atendimento médico assistencial pública abre caminho para identificar as lacunas gerenciais e econômicas

vivenciadas. Dentre elas pode-se destacar o uso clientelista e partidário dos estabelecimentos públicos, número excessivo de cargos de confiança, a burocratização das decisões e descontinuidade administrativa (TEIXEIRA, 2004). Todos esses fatores em conjunto foram provocando no longo prazo o sucateamento do SUS e a perda da sua efetividade enquanto um projeto que pretendia atender com integralidade seus usuários.

Do ponto de vista epidemiológico o SUS apresenta algumas lacunas que estão intrinsecamente relacionadas à fase da transição demográfica vivenciada pelo Brasil na época em que o modelo de tratamento foi estruturado. Estudos como de Barreto e Hage (1994; 2000) apontam para uma transição epidemiológica vivenciada pelo Brasil totalmente diferente da que fora experimentada por outros países, o que se destaca principalmente pelo ressurgimento e permanência das doenças infecto contagiosas e parasitárias.

No momento em que o perfil epidemiológico da população era em suma doenças infecciosas e parasitárias, conhecido como a primeira fase da transição epidemiológica brasileira, o SUS veio no intuito de tratar a causa do adoecimento e diminuir a proliferação do vírus (MENDEZ, 2012). Contudo, com o passar do tempo e o envelhecimento acelerado da população, o perfil das doenças agravantes havia alterado, e caminhando para as próximas fases da transição epidemiológica e demográfica, os diversos cenários indicavam para uma população que sofria de doenças crônicas degenerativas, estas que o SUS ainda não estava adaptado para atender.

Todos esses problemas econômicos e estruturais apontados resultam na falta de recursos para a realização de cirurgias, para atendimentos ambulatoriais ou para programas preventivos, abrindo uma janela de oportunidades para o setor suplementar de saúde.

2.1.2. O modelo de cuidado privado no Brasil

Antes mesmo do advento do SUS o mercado de planos de assistência à saúde no Brasil já existia, contudo, ele só veio a ser normatizado em 1998 e regulamentado após a publicação da lei 1956/98. Anteriormente a esta data, as operadoras

comercializavam seus planos sem critérios pré-definidos, baseados na ótica do mercado.

Após o advento do SUS, o setor de planos de saúde começou a ser definido como suplementar no Brasil, uma vez que se entendia que era uma opção do indivíduo pagar por um seguro privado para ter acesso à assistência médica. Sendo que a opção do usuário ter ou não um plano de saúde privado não representa um impedimento da sua utilização dos serviços assistenciais do SUS (BAHIA, 2001). Contudo, a grande discussão vivenciada neste momento era a efetiva característica suplementar do setor de saúde privado, quando se considera a existência e limitação do atendimento do serviço público, sendo o sistema privado muito mais um complemento a cobertura de determinados serviços oferecidos no sistema público.

Segundo Acioli (2006) os subsistemas público e privado irão se dividir de forma que a estruturação do SUS gera um aumento nas demandas e reorganiza a oferta dos serviços de saúde, sendo esse movimento de grande impacto na crise estrutural e econômica vivenciada naquela época pelo Brasil.

Com o processo de regulamentação do setor, ganhos para os beneficiários foram gerados, e conseqüentemente, provocou-se um aumento nos custos assistências das operadoras. De acordo com Leal (2007), o mercado de saúde suplementar enfrenta um intenso dilema regulatório entre dois objetivos: o aumento da proteção aos consumidores e a eficiência do mercado. A visão geral do legislador tem sido a de que o primeiro objetivo justifica a sua atuação como fiscalizador deste setor.

O descompasso entre as informações que o plano tem sobre os usuários e o real risco que eles apresentam para o sistema acabou por provocar desarmonia entre os dados e a assistência prestada. A literatura de economia da saúde reporta as denominadas assimetrias informacionais, sendo esta a causa das distorções que afetam o mercado de saúde suplementar de modo que este não opere com mesmo grau de eficiência que os mercados em concorrência perfeita (LEAL, 2007). Devido a estas desarmonias, muitas operadoras acabam por fechar as portas ou elevando em muito suas mensalidades.

Faz se necessário entender que a assimetria da informação não vem somente no passo da regulamentação, que voltava seus olhares para o beneficiário, mas uma parcela era destina a falta de diálogo entre os planos de assistência médico hospitalar

e o seu beneficiário, comprometendo a concretização do direito social à saúde. Segundo Bôas (2016), quando se trata da relação paciente e profissional de saúde, o compromisso de diagnosticar e de evidenciar meio para resolução da problemática vivenciada pelo beneficiário é do profissional, que acompanha o beneficiário e evidencia meios para tratar a sua doença. Contudo, a intensa interferência dos planos de saúde, diante de negativas de autorizações em intervenções cirúrgicas e procedimentos ambulatoriais, compromete a qualidade do atendimento do beneficiário e enseja riscos, tanto assistências como judiciais.

Visto a problemática, um caminho encontrado para contornar essas assimetrias estava em avançar no passo da individualização do tratamento do beneficiário, identificar precocemente as suas necessidades e tratá-las de forma antecipada. Como afirma Alves (2007), um mercado que opere sem eficiência não terá condições de suprir as demandas da sociedade em termos de preço e qualidade de planos de saúde oferecidos.

2.1.3. Interfaces da promoção e prevenção a saúde nas esferas públicas e privadas

Frente aos problemas da transição epidemiológica vivenciada pelo SUS, à adequação do tratamento as novas doenças e à regulamentação do sistema privado de assistência médica hospitalar, a utilização dos programas de prevenção e promoção da saúde se mostrou fator de eficiência na redução dos custos e controles epidemiológicos.

Para Heidmann (HEIDMANN, et al. 2006) a promoção a saúde se desponta como uma nova “concepção de saúde”. Essa nova ótica, olha não mais para o coletivo e identifica a necessidade da massa, mas torna o atendimento individualizado, identificando que cada paciente apresenta algum tipo de queixa e que cada queixa resulta de uma ação específica.

O autor vai trazer estratégias para a promoção a saúde, baseado na Carta de Ottawa, carta que reafirma a importância da promoção à saúde e aponta principalmente a influência dos aspectos sociais sobre a saúde dos indivíduos. Descrevem-se as estratégias propostas, sendo:

- 1) Implementação de políticas públicas saudáveis: não incluir apenas o tratamento, mas tudo que permeia a vida do indivíduo, como renda, meio ambiente, trabalho e outras questões;
- 2) Criação de ambientes favoráveis à saúde: é proposto a proteção do meio ambiente e a conservação dos recursos naturais, levantando como parte importante da construção da promoção a saúde;
- 3) Reorientação dos serviços de saúde: é proposta uma mudança de percurso, na qual o enfoque do tratamento passa a ser na saúde e não mais na doença;
- 4) Reforçado as ações comunitárias: são propostas ações em grupo, na intenção de reforçar a autoajuda e o apoio social;
- 5) Desenvolvimento de habilidades sociais: capacitar as pessoas a mudarem seus hábitos para uma vida saudável em que tratar da saúde se torne um hábito rotineiro.

Sustentado sob esses pilares, tanto a esfera pública como a privada desenvolveram programas de promoção e prevenção a saúde, no intuito de melhorar as condições de vida e aumento de longevidade da população brasileira, além de buscar diminuir os altos custos com tratamento de determinadas doenças e equilibrar a balança entre a receita e a despesa.

2.2. Receitas e despesas, gestões de cuidado eficientes na contenção dos custos

Os planos de saúde privado vivem sob uma balança, em que se espera que o sistema esteja em equilíbrio econômico-financeiro para que a prestação de assistência medica possa ocorrer da forma desejada. Nesta balança de um lado estão as receitas arrecadadas pelo pagamento das contraprestações pecuniárias dos beneficiários ativos e do outro os custos médico hospitalares, oriundos do tratamento em ambulatório e hospital.

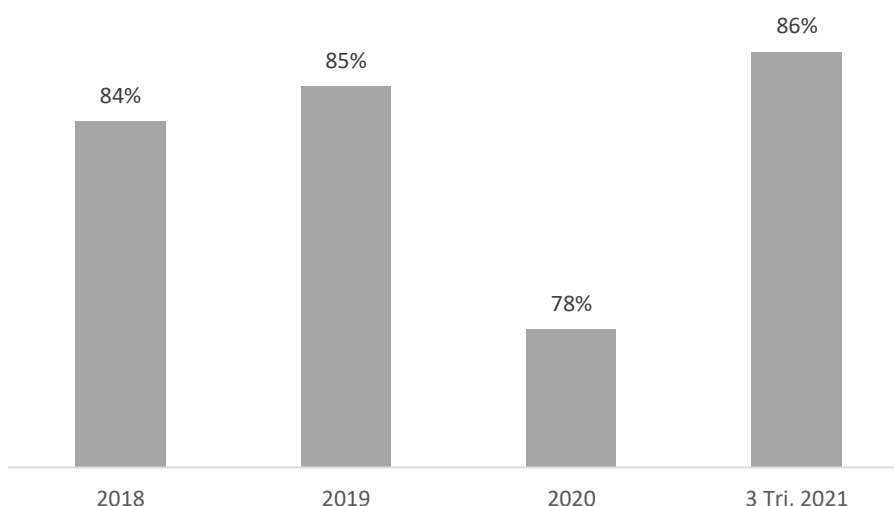
Para verificar o estado desta balança, o mercado adotou uma série de indicadores econômico-financeiros, a fim de constatar a situação de equilíbrio ou desequilíbrio vivenciada pelas operadoras de planos de saúde, entre eles, um comumente utilizado é a sinistralidade. Segundo Pires (2008), a sinistralidade é um

índice calculado através da razão entre sinistros realizados (custos de assistência) e o prêmio (receitas da assistência), medida em percentual.

Ou seja, através do indicador é possível observar o quanto do faturamento das operadoras está sendo destinado para cobrir com os gastos dos usuários. Sendo assim, operadoras com uma sinistralidade acima de 100% estão em situações de desequilíbrio econômico, uma vez que as receitas já não são suficientes para pagar os gastos assistenciais da operadora. A apuração dos últimos 12 meses das receitas versus das despesas indicam se o contrato está financeiramente estável para as partes ou se o valor pago na mensalidade é coerente para manter a relação contratual em equilíbrio, considerando os gastos assistenciais desse contrato (SILVA, 2014).

Segundo dados do Prisma Econômico Financeiro da ANS, as sinistralidades vêm aumentando de forma considerável nos últimos anos, conforme aponta o Gráfico 1. Trimestralmente é observado um aumento de 3,59% na sinistralidade das operadoras médico-hospitalares e anualmente é identificado um aumento de cerca de 11,44%.

Gráfico 1 – Sinistralidade das Operadoras Médico Hospitalares (2018 a 3º Tri 2021)



Fonte: Prisma Econômico-Financeiro da Saúde Suplementar

Com relação aos dados, vale lembrar que a queda de 2020 está profundamente relacionada a pandemia do COVID-19 que atingiu o Brasil. O isolamento social provocou a redução e em alguns casos até a paralisação dos procedimentos eletivos, impactando diretamente na redução dos custos e, por consequência, da sinistralidade.

Contudo, observa-se o indicador retornar em 2021 aos patamares observados antes da pandemia.

2.2.1. Despesas a serem honradas pelas operadoras de planos de saúde

No momento de calcular o valor ideal das contraprestações pecuniárias, é levado em conta o histórico de utilização dos beneficiários e os custos com os mesmos. As despesas das operadoras são divididas entre assistenciais e não assistenciais. De acordo com a definição da ANS em seu painel de precificação, as despesas assistenciais são aquelas destinadas ao pagamento de exames, consultas, ou seja, tudo aquilo que se refere ao tratamento do paciente em ambulatório ou hospital. Já as despesas não assistências, se referem aos gastos fixos de funcionamento das operadoras, salários, impostos, despesas com comercialização, lucro, dentre outros.

Esses gastos vêm aumentando ao longo do tempo, exigindo das operadoras cada vez melhor análise e controle dos custos assistenciais e não assistências, para que não entrem em desequilíbrio econômico-financeiro. De acordo com o apresentado no painel de precificação da ANS, com informações retiradas do Sistema de Informações de Produtos (SIP), como mostra a Tabela 1, as despesas vêm crescendo de forma rápida ao longo dos anos. Sendo que de 2019 para 2020, as despesas assistenciais cresceram cerca de 11,62%.

Tabela 1 – Despesas por ano e grupo de procedimentos (2018, 2019 e 2020)

Grupo de Procedimentos	Despesa (R\$)			Despesa a%a	
	2018	2019	2020*	2019	2020*
Consultas Médicas	25,32 Bilhões	25,77 Bilhões	20,57 Bilhões	1,8%	-20,2%
Consultas Médicas em Pronto Socorro	6,44 Bilhões	6,43 Bilhões	4,8 Bilhões	-0,2%	-25,3%
Exames	33,56 Bilhões	35,98 Bilhões	32,09 Bilhões	7,2%	-10,8%
Terapias	12,78 Bilhões	14,58 Bilhões	14,44 Bilhões	14,1%	-1,0%
Outros Atendimentos Ambulatoriais	13,29 Bilhões	14,7 Bilhões	14,22 Bilhões	10,6%	-3,3%
Demais Atendimentos Ambulatoriais	6,93 Bilhões	8,03 Bilhões	7,95 Bilhões	15,9%	-1,0%
Internações	68,17 Bilhões	80,36 Bilhões	75,59 Bilhões	17,9%	-5,9%

* Queda nos custos assistências por se tratar de um ano pandêmico e ter resultados atípicos dos habituais.

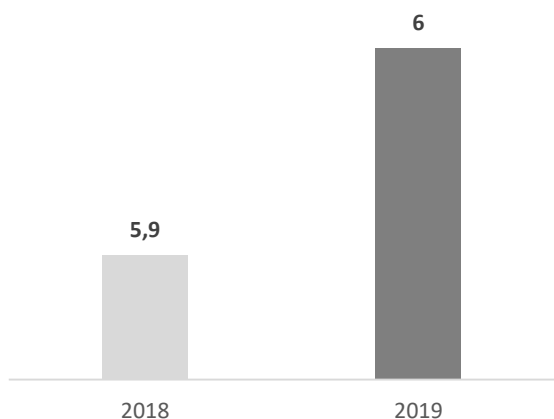
Fonte: Sistema de Informação de Produtos (SIP)

Segundo Leal (2009), para a análise da evolução dos custos assistenciais em saúde, uma relevante diferenciação é a separação em seus dois componentes, a variação dos custos médios e a variação das quantidades. A variação dos custos médios representa o aumento dos preços dos insumos (proxy da inflação) e o aumento da incorporação tecnológica cumulativa do setor. A variação das quantidades (frequência de utilização), por sua vez, pode ser associada aos argumentos da ampliação do cuidado com a saúde e do envelhecimento da população.

Segundo estudo apresentado pelo Instituto de Estudos de Saúde Suplementar – IESS, o índice de variação dos custos médicos hospitalares cresceu cerca de 14,5% de 2018 para 2019. O número expressivo apresentado pela variação dos custos para tratamento dos beneficiários, impacta diretamente no gerenciamento econômico-financeiro das operadoras e na mensalidade final que será cobrada do beneficiário.

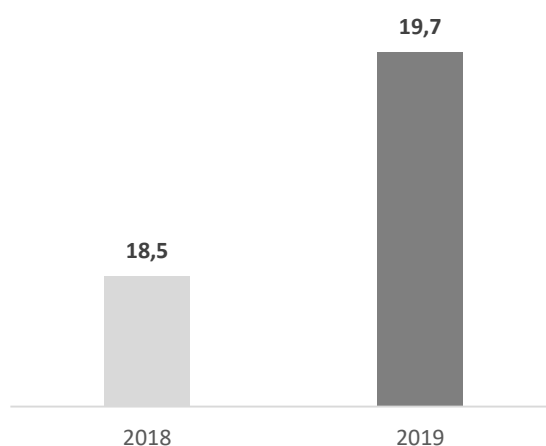
Quanto as utilizações, os planos também vêm apresentando aumento ao longo do tempo. Segundo Duarte (2016) as evoluções na utilização dos serviços de saúde impactam diretamente, assim como o gênero e a idade no custo da operadora. Observando os anos de 2018 e 2019, ignorando os resultados de 2020 por ser um ano de resultados atípicos para o setor, observa-se um aumento de cerca de 2% em consultas, de 6% na quantidade de exames e 7% no número de internações.

Gráfico 2 – Frequência de utilização em Consultas por beneficiário (2018 e 2019)



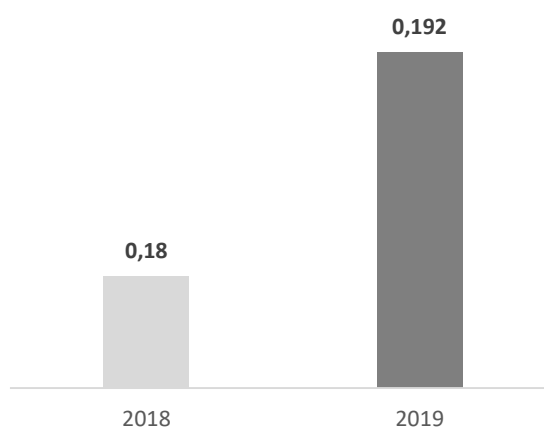
Fonte: Painel de Precificação - ANS

Gráfico 3 – Frequência de utilização em Exames por beneficiário (2018 e 2019)



Fonte: Painel de Precificação - ANS

Gráfico 4 – Frequência de utilização em Internação por beneficiário (2018 e 2019)



Fonte: Painel de Precificação - ANS

Levando em conta todas estas métricas de aumento dos custos, elevação no número de utilização, regulação prudencial, controle dos reajustes, as operadoras vêm passando por desafios cada vez maiores. Equilibrando a balança das receitas pelas despesas, a oportunidade de agir de forma preventiva e diminuindo gastos que podem ser evitados é essencial numa gestão de riscos, despertando o mercado em geral para trazer o foco para o tratamento da saúde dos usuários e não da doença.

2.3. *Outliers* na saúde suplementar: modelos para enfrentar os altos gastos dos beneficiários de alto custo

Um dos pontos que justificam as altas despesas das operadoras, são os conhecidos beneficiários de alto custo, ou também comumente conhecidos na literatura da estatística como *outliers*. De acordo com Hoppen (HOPPEN et al, 2017), os *outliers* são dados que se diferenciam drasticamente de todos os outros, ou seja, um valor que foge da normalidade e que pode causar anormalidades nos resultados obtidos por meio de algoritmos e em análises.

Esses beneficiários que fogem da curva de normalidade dos custos que em geral se espera, são frutos das assimetrias do mercado de saúde suplementar. A assimetria das informações implica que as vendas ocorram sem se ter o verdadeiro conhecimento e precisão do risco de cada comprador (LIMA, 2006). Vale ressaltar que a regulação do setor, foi de suma importância para intensificar as assimetrias informacionais, uma vez que de acordo com a Resolução Normativa 195 de 2009, é expressamente proibido na hora da contratação do plano de assistência à saúde a seleção de riscos por parte da operadora.

No âmbito da saúde suplementar, é possível destacar pelo menos três problemas que são resultantes da existência de assimetrias de informação: risco moral, indução de demanda e a seleção adversa.

Segundo Leal (2007), o risco moral ocorre quando, na presença de um plano de assistência médica, a estrutura de incentivos que o usuário se depara se altera, favorecendo a utilização de serviços além do limite em que ele utilizaria caso pagasse diretamente por aquele serviço.

Já a assimetria relacionada a indução da demanda, é refletida pela responsabilidade pelo diagnóstico e pelo tratamento que é uma tarefa delegada ao médico. Neste contexto, o corpo clínico são os responsáveis por direcionar o paciente dentro do setor médico hospitalar, surgindo a possibilidade de criação de demanda pelos seus próprios serviços, gerando assimetria entre o paciente e o médico, e entre o médico e a operadora.

Por fim, Leal aponta para a seleção adversa, fenômeno que ocorre no caso em que a operadora não é capaz de conhecer perfeitamente o risco dos indivíduos (ou a probabilidade de adoecimento) antes de aceita-los em suas carteiras. Frente a estas assimetrias, o mercado vem propondo uma série de ações corretivas, com foco principal na prevenção e tratamento precoce, como os Programas de Promoção e Prevenção a Saúde, implantados pela ANS.

Porém, estas ações ainda não são suficientes para gerar uma diminuição do impacto causado pela seleção adversa, como o aumento dos custos. Pelo contrário, o evento que é observado comumente pelas operadoras é que indivíduos com risco inferior ao risco médio não aderem ao contrato, pois seria caro demais para eles, e no final do processo apenas indivíduos com alto grau de risco permanecem no plano, o que inviabiliza a existência do mercado privado de saúde.

2.3.1. Beneficiários de alto custo nas carteiras dos planos de saúde

Segundo a grande parte da literatura, os beneficiários de alto custo são aqueles que representam riscos mais elevados e custo de tratamento acima da média de toda a carteira. Segundo Blumenthal (2016), os beneficiários que são mais onerosos para a carteira possuem condições médicas mais graves e são acometidos muitas das

vezes por eventos adversos previsíveis, por estarem em maior contato com o sistema de saúde em relação aos demais beneficiários.

Além disso, os beneficiários de tratamentos mais caros possuem forte correlação com o sexo e a idade, no geral, o alto custo se concentra nas faixas mais idosas, pois os idosos possuem necessidades e tratamentos específicos que estão intrinsecamente relacionados a idade (VERAS et al., 2014).

. Segundo estudo apresentado pelo IESS (2015): “Caracterização dos beneficiários de alto custo assistencial - Um estudo de caso”, é esperado um aumento no número de beneficiários dispendiosos com o envelhecimento da população.

Em termos percentuais, esses beneficiários representam uma pequena parcela da carteira, mas, em relação aos gastos assistências, eles podem representar grande porcentagem do total. Ainda em relação ao estudo realizado pelo IESS, em 2015, para uma autogestão, foi observado que os 5% de beneficiários que geraram maior despesa assistencial foram também responsáveis por 66,5% das despesas assistências totais da operadora. Ainda que as autogestões apresentem um padrão de utilização diferente do restante do setor, o resultado serve de baliza em relação aos custos desses usuários para as outras operadoras, já que o padrão de utilização e a procura de tratamentos sofisticados é a mesma.

Em relação ao perfil das doenças, os 5% dos beneficiários mais custosos para os planos de saúde adoecem por problemas crônicos, em geral. A correlação entre o perfil de morbimortalidade desses pacientes e os custos dos beneficiários está ligado a alguns fatores como: (i) o complexo movimento de coordenação do cuidado entre os vários prestadores de assistência à saúde, além da duplicação de exames e procedimentos, (ii) o aumento e intensivo uso de especialistas ou (iii) a maior probabilidade de internações (ALDRIDGE et al., 2015).

Frente a esse cenário, as operadoras buscam meios para solucionar a problemática, que entra em conflito com a não seleção de riscos, o aumento progressivo nos custos assistenciais e a intensificação da seleção adversa. A alta representatividade dos beneficiários de alto custo na despesa assistencial torna de suma importância entender as características que definem esse grupo e o porquê de gerarem despesas tão elevadas.

2.3.2. Adoção de modelos preditivos no mercado de saúde suplementar

Após identificado o problema da falta de previsibilidade quanto ao risco que o beneficiário representa para a operadora, muitas delas começaram a adotar medidas de prevenção a riscos e doenças, a fim de tratar de forma antecipada o problema do beneficiário e reduzir os custos assistenciais que ele no futuro poderia gerar.

O termo 'prevenir' tem o significado de "preparar; chegar antes de; dispor de maneira que evite (dano, mal); impedir que se realize" (FERREIRA, 1986). As ações preventivas das operadoras vêm sendo amplamente discutidas, uma vez que o mercado ainda está bem atrás em relação as pesquisas sobre o tema.

A literatura já vem apontando o quanto a adoção de modelos para identificação de doenças e riscos de forma antecipada e tratamento precoce reduz os custos operacionais. Mas, o desafio dos modelos de predição com o objetivo de prevenção vai além de categorizar os fatores de risco, dando a eles seus devidos pesos e medidas. Consiste também em ajustar o foco, muitas vezes distorcido, do risco percebido versus o risco real (CARVALHO, 2013).

Segundo Júnior (JÚNIOR, et al, 2020), o acesso a informações e a forma como se trabalha com ela se tornam um grande diferencial dentro do mercado. Isso possibilita às operadoras observar sua carteira e trazer um olhar individualizado para a problemática vivenciada por cada um de seus beneficiários, trazendo a oportunidade de adiantar o tratamento e evitar o adoecimento ou agravamento de diversas doenças.

Dessa forma, destacam-se os modelos preditivos. Esses modelos são baseados em dados históricos das empresas, de onde é possível extrair informações que mostram o real estado da natureza do problema, ou seja, os dados extraídos do meio em que o plano está exposto mostram a natureza do problema em questão (LOPES, 2018).

Leva-se em conta os eventos ocorridos no passado, com o auxílio de ferramentas estatísticas, com o objetivo de prever com certo nível de acurácia, comportamentos futuros (LENTZ, 2013). No entanto, as dificuldades encontradas para se construir modelos de predição em Saúde se inicia no acesso e veracidade da informação, uma vez que o diagnóstico é dado pela classificação internacional das doenças (CID), informação que é muitas vezes desconsiderada pelo corpo clínico no preenchimento de laudos.

A falta do dado acaba prejudicando a modelagem preditiva, uma vez que o registro dos eventos passados, que configuram parte essencial para a modelagem preditiva, vem incompleta e incorreta, faltante ou não existe em certas bases de dados de operadoras de planos de saúde. Além de alertar o corpo médico e clínico sobre a importância do registro correto dos dados e sobre como o tratamento desta informação pode gerar redução de custos para as operadoras e resultados financeiros expressivos, os problemas de registro de dados podem ser contornados quando se utilizam bases robustas e com quantidade expressiva de informações.

Frente às problemáticas destacadas, é possível observar o quanto a regulação do setor de saúde suplementar representou em ganhos e ônus, e, por mais que os planos de saúde ganharam espaço com o sucateamento do SUS, eles tiveram uma diversidade de desafios com o avanço das tecnologias em saúde e o aumento dos custos de diversos procedimentos. Desse modo, faz-se necessária a construção de modelos que buscam individualizar a demanda de cada beneficiário e tratar de forma antecipada as suas doenças. Por mais que o tema ainda seja pouco explorado, a experiência de outras áreas pode auxiliar na construção dos modelos robustos eficazes para a área da saúde.

3. MODELOS UTILIZADOS PARA A CLASSIFICAÇÃO

3.1. Regressão Logística

Em muitos casos, os pesquisadores desejam alocar os objetos em classes, para isso é comumente utilizado métodos de classificação. Quando se deseja prevenir a classificação de certos atributos baseado em outros atributos, a regressão logística é comumente levantada como meio para realização desta classificação, utilizando-se de variáveis dicotômicas como variável resposta, representadas por exemplo, 1 ou 0 (GONZALES, 2018).

A regressão logística tem como diferencial a categorização da sua variável resposta, e mesmo quando ela não é dicotômica, é possível torná-la dicotômica. Em relação às variáveis independentes, essas podem ser de qualquer tipo.

Os modelos logísticos estimam a probabilidade de obtenção de uma das categorias da variável dependente, ou seja, a probabilidade de ocorrência de

determinado evento, a partir de uma função das variáveis preditoras incluídas no modelo.

Para Hair e colegas (1998), existem uma série de fatores que estimulam a utilização da regressão logística nas análises de classificação, dentre elas:

- É uma técnica mais genérica e robusta, pois sua aplicação é apropriada em grande variedade de situações;
- É uma técnica similar a regressão linear múltipla.

Neste estudo, a variável resposta utilizada foi adoecimento pelo desfecho, para a qual 0 indica o não adoecimento e 1 indica o adoecimento pelo desfecho escolhido.

No modelo logístico, a relação entre a probabilidade da ocorrência do desfecho é dada pela seguinte função sigmoide:

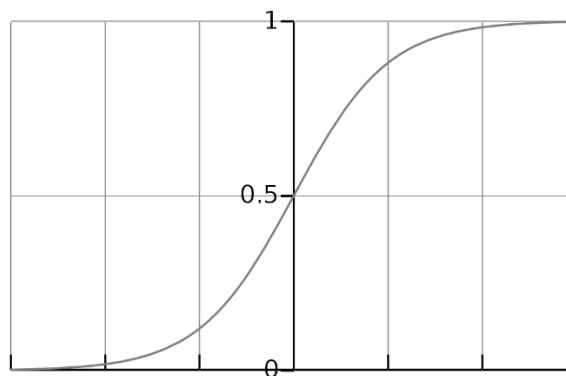
$$P(Y = 1|X) = \frac{e^{\mu}}{1 + e^{\mu}} \quad (1)$$

na qual μ é uma função que relaciona as variáveis preditoras e tem formulação:

$$\mu = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i + \epsilon \quad (2)$$

A função em (1) irá assumir valores que vão de 0 a 1 para $\mu \in (-\infty, +\infty)$.

Gráfico 5 – Função Sigmóide



Fonte: Elaboração própria

Os coeficientes de regressão $\beta_0, \beta_1, \dots, \beta_x$ são estimados a partir do conjunto de dados, pelo método da máxima verossimilhança, que encontra uma combinação de coeficientes que irá maximizar a probabilidade da amostra ter sido observada (HOSMER et al, 1989). A partir dos coeficientes de regressão estimados, é calculado o valor de μ para cada indivíduo e, a partir da função em (1), é calculada a estimativa da probabilidade de ocorrência do desfecho para cada indivíduo.

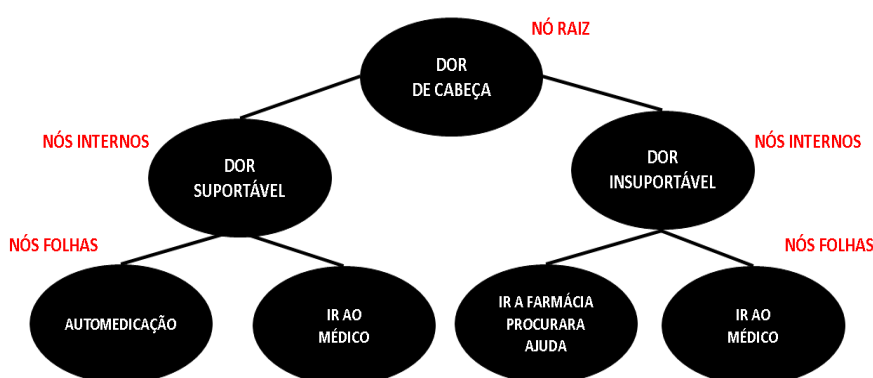
3.2. Árvore de Classificação

Outro método de classificação comumente utilizado pela literatura é a denominada árvore de classificação. Formalmente, se caracteriza como uma metodologia não paramétrica que cria uma partição no espaço das covariáveis em regiões distintas e disjuntas (LOPES, 2018).

Segundo Meira e colegas (2008), a árvore de classificação é um modelo representado por nós e ramos, parecido com uma árvore, porém no sentido invertido. O nó raiz é o primeiro nó da árvore, no topo da estrutura. Os nós internos, incluindo o nó raiz, são nós de decisão. Cada nó de decisão possui um teste sobre uma variável independente e os resultados desses testes formam os ramos da árvore. Os nós folhas, concentrados nas extremidades das árvores, representam valores de predição para as variáveis dependentes ou a distribuição de probabilidade desses valores.

A Figura 1 exemplifica a estrutura da árvore de decisão, para uma situação hipotética (pessoa ir ou não ao médico). O nó raiz é a existência de uma dor. Essa situação gera dois nós internos, nos quais a pessoa avalia se a dor é suportável ou insuportável e, por fim, os nós folhas ou ramos, nos quais é gerada a distribuição de probabilidade para a decisão final do indivíduo (ir ao médico, se automedicar ou ir até a farmácia).

Figura 1 – Estrutura da árvore de classificação



Fonte: Elaboração própria

O propósito básico de uma árvore de classificação é produzir um modelo de predição preciso e descobrir a estrutura preditiva do problema (BREIMAN et al., 1984). Desta forma, ao utilizar da árvore de classificação, o presente trabalho buscou

encontrar a estrutura de utilização dos indivíduos antes do adoecimento pelo desfecho escolhido.

Para determinar qual será a partição do nó raiz em nós internos, considera-se primeiramente o conjunto de todas as possíveis partições do conjunto amostral, denotado por $S(t)$. Cada regra de partição (s) do conjunto de possíveis partições ($S(t)$) será definida em função de uma das variáveis independentes X_j ($j = 1, 2, \dots, n$) do vetor x .

No caso de variáveis explicativas categóricas, ou seja, ir ou não, ir ter ou não ter, adoecer ou não por determinado desfecho, a regra de partição pode ser definida da seguinte forma:

$$S_1 : T_L = \{ X \in T \mid X_1 \in \{1\} \} \text{ e } T_R = \{ X \in T \mid X_1 \in \{2\} \}$$

Onde:

X_1 é uma variável categórica que assume os valores $\{1, 2\}$

T_L é nó interno que se desmembra do nó raiz para a esquerda

T_R é nó interno que se desmembra do nó raiz para a direita

Em seguida é necessário determinar a regra de partição de s e S que melhor irá separar as variáveis independentes da variável dependente. Essa regra será determinada como partição ótima e denotada como s^* . Tendo definido qual será a repartição ótima, o nó raiz é repartido em nós internos de acordo com s^* . O processo será repetido até que um nó interno se torne um nó folha, ou seja, as informações provenham da mesma população ou tenha uma quantidade de informações de outra população insignificante.

Um dos critérios para determinar qual será a partição ótima s^* é o critério de Gini, que irá se basear no somatório das variâncias das observações em um determinado nó. A i -ésima população será denotada como τ_i , $i = 1, 2, \dots, g$ e redefinindo as observações no nó raiz como 1 se a observação vier de τ_i e 0, caso não venha de τ_i .

A variância amostral será estimada por

$$\hat{P}(\tau_i|T)(1 - \hat{P}(\tau_i|T))$$

e,

$$\hat{P}(\tau_i|T) = \sum_{j=1}^g \frac{\hat{P}_j n_i(t)}{n_i \hat{P}(t)}$$

Será denotado como p_i a probabilidade a priori da i -ésima população, que será estimado por $\frac{n_i}{n}$, ou será considerada igual para todas as populações, ou seja, $\frac{1}{g}$, $n_i(t)$ é o número de casos na população i no nó raiz, n_i é o número de casos na população i e $\hat{P}(t)$ é a probabilidade estimada de que uma observação esteja no nó raiz, que será dado por

$$\hat{P}(t) = \sum_{j=1}^g \frac{\hat{P}_j n_i(t)}{n_i}$$

Ao se somar todas as variâncias para todas as populações, obtém-se a média de impureza de Gini, que será dada por

$$i(t) = \sum_{j=1}^g \hat{P}(\tau_i|T)(1 - \hat{P}(\tau_i|T))$$

Medida que irá indicar a grau de heterogeneidade das observações no nó em questão, ou seja, quanto menor o índice, mais homogêneas são as informações no nó.

A medida de Gini pode ser reescrita como

$$i(t) = 1 - \sum_{j=1}^g \hat{P}^2(\tau_i|T)$$

e,

$$\sum_{j=1}^g \hat{P}(\tau_i|T) = 1$$

A medida tomará seu valor máximo quando $\hat{P}^2(\tau_i|T) = \frac{1}{g}$ para $i = 1, \dots, g$, e o valor mínimo será dado quando o nó t for um nó folha, ou seja, quando $\hat{P}^2(\tau_i|T) = 1$, para algum i e $\hat{P}^2(\tau_i|T) = 0$, para todo $h \neq i$. Por isso, $i(t)$ é determinado como uma medida de impureza do nó raiz.

A melhor partição de um nó raiz em um nó interno é dada pela regra ótima s^* que maximiza

$$\Delta i(s, t) = i(t) - P_L i(t_L) - P_R i(t_R)$$

Em que P_L é a probabilidade de uma observação ser classificada como membro do nó interno descendente esquerdo e P_R é a probabilidade de uma observação ser classificada como membro do nó interno descendente direito. A partição ótima será aquela que maximiza o decréscimo na impureza da árvore inicial $i(t)$.

A função $\Delta i(s, t)$ será maximizada em dois passos, sendo eles:

- Passo 1 – Para as variáveis independentes qualitativas, busca-se a categoria ou o subconjunto “A” que maximiza $\Delta i(s, t)$, ou seja, se a observação do nó raiz pertence a categoria A, então ela é alocada no nó interno descendente à esquerda, caso contrário, é alocado no nó interno descendente à direita. Se a variável independente é quantitativa, encontra-se um valor c que maximiza $\Delta i(s, t)$, ou seja, se uma observação é menor ou igual a c , ela será alocada no nó interno descendente à esquerda, caso contrário, é alocado no nó interno descendente à direita.
- Passo 2 – É escolhido a variável que obtém o maior valor de $\Delta i(s, t)$ em relação a todas as variáveis independentes. O processo de partição segue nos nós internos descendentes até que a árvore não possa ser mais particionada, ou seja, quando atinge o nó folha. Esta árvore que chegou no estágio do nó folha é denominada árvore máxima ($T_{\text{máx}}$).

A árvore máxima obtida pode não produzir boas regras de predição, pois depende da amostra utilizada. Para evitar a ocorrência deste fenômeno, denominado como *overfitting*, o próximo passo é encontrar a árvore ótima, aquela que possui o tamanho correto T^* , através do corte (poda) da árvore $T_{\text{máx}}$.

Para determinar a árvore ótima é construído uma sequência de árvores $T_0 = T_{\text{máx}}, T_1, T_2, \dots, T_n = t_1$ candidatas a serem T^* . Partindo-se de $T_{\text{máx}}$, o objetivo é definir como as podas devem ser feitas até chegar em t_1 . Uma vez que se obtém as candidatas a T^* , escolhe-se como árvore ótima aquela que minimiza a taxa de erro real, que usualmente refere-se a taxa de erro da amostra teste.

Na literatura de *machine learning*, os pesquisadores buscam se precaver na hora da montagem dos modelos para que não ocorrer fenômenos relacionados ao sobreajuste, ou conhecido como *overfitting*. O fenômeno ocorre quando nos dados de treino o modelo possui uma performance excelente, porém, quando o modelo é usado em dados de teste, o resultado é ruim.

O sobreajuste acontece porque o modelo aprendeu peculiaridades no treino e acabou provocando uma série de padrões que deveriam ser reproduzidos, e quando o modelo recebe as informações das variáveis preditoras nos dados de teste, o modelo aplica as mesmas métricas aprendidas. Porém, pelo fato de os dados serem diferentes, as regras aprendidas não possuem validade.

Na intenção de corrigir essas generalizações, é necessário fazer boas seleções de variáveis e aplicação correta dos modelos de classificação.

4. METODOLOGIA

Nesta seção, serão descritos a construção do banco de dados utilizado e o procedimento para a análise desses dados. Todo o trabalho foi realizado utilizando o software R, que é um programa livre para análise de dados, conjuntamente com pacotes estatísticos disponíveis, como dplyr, rpart, ggplot2, randomForest, tidyverse, tidymodels, funModeling, vip, janitor, xgboost, dentre outros.

A fim de atender os objetivos propostos neste trabalho, foi realizado um estudo de caráter empírico e exploratório, de abordagem quantitativa, uma vez que esta abordagem tem caráter sistemático de coleta e análise dos dados, a fim de descrever o fenômeno em estudo (FORTIN, 2013).

O método aqui proposto utiliza-se também de uma abordagem comparativa, onde pretende-se a partir da definição do problema, comparar dois métodos em relação ao seu poder preditivo.

4.1. Base de dados

Os dados utilizados neste trabalho foram provenientes do Sistema de Informação Ambulatorial do SUS (SIA-SUS) e do Sistema de Informação Hospitalar do SUS (SIH-SUS), e mais especificamente, os bancos de Laudos Diversos, retirados do conjunto de dados do SIA-SUS, e os bancos RD*, retirados do conjunto de dados do SIH-SUS.

Para realização das análises exploratórias dos dados e conseqüentemente propor o modelo preditivo, foi realizado um recorte temporal na base de onze anos, sendo as informações referentes de janeiro de 2008 até dezembro de 2018.

Os bancos de Laudos Diversos têm informações sobre o atendimento ambulatorial realizado em pacientes do Sistema Único de Saúde. A base possui 45 variáveis, que estão divididas em variáveis sobre o paciente como: sexo, idade, etnia, cor, local onde mora; variáveis relacionadas ao tipo de atendimento (tipo de estabelecimento em que foi atendido, grau de risco do atendimento, data do processamento, tipo de prestador); e variáveis sobre os procedimentos e diagnósticos dos pacientes. A base abriga as informações de todos os estados brasileiros e ao total, considerando todo o recorte temporal, possui 6,57 bilhões de registros de procedimentos ambulatoriais e hospitalares realizados no SUS.

Os bancos RD* tem informações sobre o atendimento hospitalar realizado em pacientes do Sistema Único de Saúde, com informações sobre cirurgias, internações de emergência ou eletivas. A base possui 114 variáveis, que estão divididas em variáveis sobre o paciente como sexo, idade, etnia, cor, local onde mora, variáveis relacionadas à internação do paciente (tipo de estabelecimento em que foi atendido, grau de risco do atendimento, data do processamento, tipo de prestador, caráter da internação, utilização de leito ou especialidade, dentre outras informações) e variáveis sobre os procedimentos e diagnósticos dos pacientes. A base abriga as informações de todos os estados brasileiros e ao total, considerando todo o recorte temporal, possui 140.612.096 milhões de registros de procedimentos hospitalares realizados no SUS.

A justificativa de utilização deste banco de dados é a sua robustez, o que provoca maior número de informações para análise e confiabilidade para os resultados encontrados no modelo. Faz-se necessário esclarecer que os modelos preditivos têm o foco em estimar a probabilidade de adoecimento dos pacientes para certa doença, sendo o tratamento da enfermidade parecido tanto no sistema público, na qual se está utilizando, como no sistema privado. Desta forma o modelo pode ser utilizado para predição em ambos os cenários, público e privado.

4.2. Análise descritiva dos dados brutos

A fim de identificar qual era o perfil da população a ser estudada e quais eram as doenças com maiores custos de tratamento dentro do SUS, foram realizadas

análises descritivas dos dados de ambas as bases analisadas, de Laudos Diversos do SIA-SUS e a RD* do SIH-SUS.

As análises foram realizadas de acordo com os grupos de variáveis apresentadas no tópico anterior, variáveis com as características dos beneficiários, sobre o tipo atendimento e variáveis sobre os procedimentos realizados e diagnóstico dos usuários.

Foram realizadas análises exploratórias quantitativas e análises gráficas, além da criação de ranqueamento das informações para as variáveis de grupo de procedimentos e variáveis CID por capítulo e categoria, em que se ranqueou as informações por quantidade total de utilização, custo total das utilizações e custo médio.

Após finalizado o descritivo do banco de dados, avaliou-se quais doenças eram mais prevalentes na população e mais onerosas em seu tratamento. Foi feito um recorte, em que se considerou uma doença de alto custo de tratamento aquelas que possuíam mais de mil reais pelo custo médio por procedimento.

Dentre as mais de quatro mil doenças listadas na Classificação Internacional de Doenças (CID), uma parcela de quase quatrocentas doenças foram classificadas como alto custo pelo recorte acima citado. Destas, foi escolhida uma doença que tivesse incidência tanto em homens como em mulheres e que abrangesse qualquer idade, não estando concentrada na velhice, infância, fase adulta ou juventude. Além disso, foi escolhida a doença que apresentasse prevalência não muito baixa na população, ou seja, com quantidade relevante de registros dentro do recorte temporal apresentado.

4.3. Pré-processamento das bases de dados

Para realizar a modelagem preditiva, foi realizada uma preparação das bases de dados de Laudos Diversos e RD*. Logo em seguida, foram selecionadas apenas variáveis que seriam importantes na montagem do modelo. O Quadro 1 descreve as variáveis que foram selecionadas de ambos os bancos.

Quadro 1 – Descrição das variáveis selecionadas

Variável	Descrição
COD_PAC	Código do Usuário. Os usuários quando realizam seu primeiro atendimento no SUS, recebem um código de acompanhamento. Em toda a sua trajetória no SUS, o código será o mesmo para o usuário. Este código é criptografado, não sendo possível identificar nome ou dados pessoais do usuário.
Data de Processamento	Data de ocorrência do procedimento realizado pelo usuário dentro do SUS
UF	Estado em que o usuário mora no momento em que realizou o procedimento
Gênero	Sexo do Usuário (Masculino/Feminino)
Idade	Idade do usuário na época em que realizou o procedimento
Procedimento	Código de 10 dígitos que faz referência ao procedimento realizado pelo usuário
CID	CID por categoria referente ao tipo de procedimento realizado pelo usuário

Fonte: Elaboração própria

A variável de interesse neste trabalho foi a ocorrência do nosso desfecho, que estava na variável CID. As demais variáveis foram suporte no tratamento das bases para que fosse possível identificar os beneficiários que tiveram o desfecho e selecionar a sua utilização durante todo o período.

Vale ressaltar, que da nossa base tratada, a única variável que podia indicar diagnóstico do usuário era o CID, pois o mesmo é uma ferramenta epidemiológica utilizada para classificar o estado de saúde dos usuários pelo médico.

4.4. Identificação e histórico do usuário

Após processadas as bases, foi realizada a coleta das informações dos beneficiários que tinham o registro do desfecho escolhido anteriormente, a fim de verificar qual é o padrão de utilização desses beneficiários antes de adoecerem.

Desta forma, para avaliar qual o padrão de procedimentos realizados no SUS pelos indivíduos antes da ocorrência da doença, é introduzido dois conceitos: período de avaliação e período de corte. O período de avaliação se refere ao período de dois anos antes de aparecer o primeiro registro CID da doença escolhida. No período de

avaliação, foram registrados os procedimentos utilizados pelos usuários. Sendo assim, foram expurgados desta base todos os usuários que tiveram o desfecho no período de avaliação, pois caso esses permaneçam, o modelo pode ficar enviesado. Já o período de corte foi a competência referente a um mês após o período de avaliação, no qual identificou-se os usuários que tiveram o desfecho e os que não o tiveram.

Para identificar os usuários segundo ter ou não o desfecho, foi criada uma variável indicadora, que valeu 1 para os usuários que tiveram o desfecho no último mês e 0 para os que não tiveram. Ao final a base contou com informações de todos os indivíduos que tiveram ou não a doença no período de corte, mas que não tiveram o desfecho no período de avaliação.

De forma hipotética, a Figura 2 apresenta uma base que tem como período de avaliação: janeiro/2008 até dezembro/2019, e período de corte: janeiro/2010.

Figura 2 – Base hipotética para análise

	COD_PAC	DATA	UF	GÊNERO	IDADE	PROCEDIMENTO	CID
PERÍODO DE AVALIAÇÃO	A	01/2008	MG	M	32	501030018	D43
	B	01/2008	ES	F	18	304070050	C14
	C	03/2008	RO	M	52	501030026	H15
	D	05/2008	TO	F	22	304070068	H62
	A	08/2008	MG	M	32	304070050	H14
	B	04/2009	SP	F	19	701020474	B23
	A	06/2009	MG	M	33	501030026	C99
	C	09/2009	RO	M	53	701020458	D190
	D	09/2009	TO	F	23	701020474	G19
PERÍODO DE CORTE	A	01/2010	MG	M	34	701020458	H62
	E	01/2010	GO	F	65	304070050	D32

Fonte: Elaboração própria

Selecionados os períodos, foram expurgados da base os usuários que tiveram o desfecho escolhido (para fins ilustrativos, CID “H62”) no período de avaliação, ou seja, o usuário D. A Figura 3 ilustra essa operação.

Figura 3 – Limpeza da base de análise

		COD_PAC	DATA	UF	GÊNERO	IDADE	PROCEDIMENTO	CID
PERÍODO DE AVALIAÇÃO	A	A	01/2008	MG	M	32	501030018	D43
	B	B	01/2008	ES	F	18	304070050	C14
	C	C	03/2008	RO	M	52	501030026	H15
	D	D	05/2008	TO	F	22	304070068	H62
	A	A	08/2008	MG	M	32	304070050	H14
	B	B	04/2009	SP	F	19	701020474	B23
	A	A	06/2009	MG	M	33	501030026	C99
	C	C	09/2009	RO	M	53	701020458	D190
	D	D	09/2009	TO	F	23	701020474	G19
	PERÍODO DE CORTE	A	A	01/2010	MG	M	34	701020458
E		E	01/2010	GO	F	65	304070050	D32

↓

		COD_PAC	DATA	UF	GÊNERO	IDADE	PROCEDIMENTO	CID	
PERÍODO DE AVALIAÇÃO	A	A	01/2008	MG	M	32	501030018	D43	
	B	B	01/2008	ES	F	18	304070050	C14	
	C	C	03/2008	RO	M	52	501030026	H15	
	A	A	08/2008	MG	M	32	304070050	H14	
	B	B	04/2009	SP	F	19	701020474	B23	
	A	A	06/2009	MG	M	33	501030026	C99	
	C	C	09/2009	RO	M	53	701020458	D190	
	PERÍODO DE CORTE	A	A	01/2010	MG	M	34	701020458	H62
		E	E	01/2010	GO	F	65	304070050	D32

Fonte: Elaboração própria

Logo, como o usuário “D”, apresentou registro de “H62” dentro do período de avaliação, todos os seus registros independentes do CID ser ou não “H62” foram excluídos. Essa tratativa é feita para que o modelo não identifique um padrão de um beneficiário que já possui a doença, enviesando a análise.

Feita a limpeza da base de dados, foi criada uma coluna “Indicador”, que atribuiu 1 para os beneficiários que tiveram o desfecho no período de corte e 0 para os beneficiários que não tiveram, conforme ilustrado na Figura 4.

Figura 4 – Criação da variável Indicador na base de análise

	COD_PAC	DATA	UF	GÊNERO	IDADE	PROCEDIMENTO	CID	INDICADOR
PERÍODO DE AVALIAÇÃO	A	01/2008	MG	M	32	501030018	D43	1
	B	01/2008	ES	F	18	304070050	C14	0
	C	03/2008	RO	M	52	501030026	H15	0
	A	08/2008	MG	M	32	304070050	H14	1
	B	04/2009	SP	F	19	701020474	B23	0
PERÍODO DE CORTE	A	06/2009	MG	M	33	501030026	C99	1
	C	09/2009	RO	M	53	701020458	D190	0
	A	01/2010	MG	M	34	701020458	H62	1
	E	01/2010	GO	F	65	304070050	D32	0

Fonte: Elaboração própria

Sendo assim, como o beneficiário “A” teve o registro de “H62” no período de corte, todos os seus registros no período de avaliação receberam 1. Os demais beneficiários: “B”, “C” e “E”, como não registraram o desfecho selecionado, receberam 0 em todos os seus registros no período de avaliação.

Este processo foi realizado em todo o recorte temporal da base de dados de 2008 até 2018, sendo que o período de corte se iniciou em janeiro de 2010 e foi até dezembro de 2018 e o de avaliação foi de janeiro de 2008 até novembro de 2018.

Finalizada a limpeza e criação da coluna indicador, foi realizada uma descrição das bases geradas, segundo o sexo, a idade, a UF dos usuários, e o mais importante, segundo os procedimentos realizados e CIDs registrados para esses usuários no período de avaliação.

Nesta análise, a ideia foi criar uma escada que mostrasse qual é caminho percorrido por cada usuário até adoecer pelo desfecho escolhido, além de entender quais os procedimentos, seja exames, consultas, terapias, normalmente realizados pelos usuários que tiveram o desfecho estudado.

4.5. Montagem da base final

Antes de modelar os dados, foi realizado um último processamento das bases. Nos dados anteriormente gerados, os registros dos beneficiários poderiam ser encontrados em diversas linhas, já que cada procedimento realizado gerava uma nova observação.

Desta forma, a base final foi o resultado da compilação de todos os registros em um único, ou seja, foi suprimida todas as informações dos usuários em uma única linha. Para realização deste procedimento, foram criadas variáveis que continham os procedimentos e CIDs registrados pelos usuários em um período de dois anos de utilização do SUS.

As novas variáveis foram divididas em cinco grupos, conforme mostra o Quadro 2.

Quadro 2 – Grupo de variáveis criadas

Grupo de Variáveis	Descrição	Quant. de Variáveis
Beneficiário	Variáveis que fazem referência aos beneficiários (Sexo, Gênero, Idade, UF, Código do Usuário e Indicador)	5
Gerais	Variáveis que fazem a detecção de quantas vezes os beneficiários utilizaram dos serviços do SUS em dois anos, um ano e seis meses, em um ano e em seis meses.	16
Procedimento	Variáveis que fazem a detecção de quantas vezes os beneficiários realizaram certos procedimentos em dois anos, um ano e seis meses, em um ano e em seis meses.	23
CID	Variáveis que fazem a detecção de quantas vezes os beneficiários tiveram o registro de certos CIDs em dois anos, um ano e seis meses, em um ano e em seis meses.	103
Time	Variáveis que fazem a detecção de quanto tempo atrás os beneficiários fizeram determinados procedimentos ou tiveram registro de alguns CID	96

Fonte: Elaboração própria

Para a criação das variáveis gerais, levou-se em conta quantas vezes os beneficiários utilizaram os serviços do SUS por grupo de procedimentos, sendo eles descritos no Quadro 3.

Quadro 3 – Grupo de procedimentos

Código	Grupo de Procedimentos
301	Consultas / Atendimentos / Acompanhamentos
201	Coleta de material
701	Órteses, próteses e materiais especiais não relacionados ao ato cirúrgico
503	Ações relacionadas à doação de órgãos e tecidos para transplante
504	Processamento de tecidos para transplante
405	Cirurgia do aparelho da visão
505	Transplante de órgãos, tecidos e células
506	Acompanhamento e intercorrências no pré e pós-transplante
206	Diagnóstico por tomografia
307	Tratamentos odontológicos
409	Cirurgia do aparelho geniturinário
309	Terapias especializadas
211	Métodos diagnósticos em especialidades
414	Bucomaxilofacial
418	Cirurgia em nefrologia

Fonte: Elaboração própria

Ao total foram geradas 243 variáveis, que são apresentadas no Apêndice A.

A Figura 5 mostra uma parte da base final após todos os processamentos. As quatro últimas colunas ilustram as variáveis que foram criadas no banco final, sendo que QU2A representa a quantidade de utilizações dentro dos últimos 2 anos, QU2A_211020010 representa a quantidade de vezes que realizou o procedimento 211020010 dentro dos últimos 2 anos e QU2A_H62 representa a Quantidade de vezes que foi registrado o CID H62.

Figura 5 – Aspecto geral da base final

COD_PAC	INDICADOR	UF	GÊNERO	IDADE	QU2A	QU2A_211020010	QU1A_H260	QTC_Z947
A	1	MG	M	32	5	1	4	1
B	0	SP	F	18	6	0	5	9
C	0	RO	M	52	7	2	8	0

Fonte: Elaboração própria

Na base final, cada usuário apareceu apenas uma única vez. Caso ele tenha tido registros da doença no período de corte e apareça novamente nos períodos de avaliação posterior, foi sempre selecionado apenas o recorte dos dois últimos anos

mais recentes do seu histórico no SUS. Além disso, caso um beneficiário tenha tido o desfecho em um período de corte e ficou por 2 anos de avaliação sem aparecer na base, e em seguida apareceu, foi selecionado o intervalo de tempo em que houve a maior quantidade de registros de utilização.

Quanto às variáveis relacionadas aos usuários, como local onde mora e idade, foi selecionada a informação mais recente, ou seja, as informações referentes ao seu último registro dentro do período de avaliação.

Após finalizado o processamento das bases finais, foi realizada uma última descrição dos dados, em que foram comparadas as distribuições de frequências de todos os usuários com a distribuição entre os usuários que tiveram o desfecho. Esta avaliação teve o objetivo de identificar as frequências de cada uma das variáveis geradas e expurgar da base as variáveis que possuem pouca variabilidade, para que não ocorra *Overfitting* durante o processamento do modelo.

4.6. Ambiente Computacional

Após a estruturação da base final, foram verificadas as propriedades de cada uma das variáveis. Foram observados a quantidade de zeros e “NAs” em cada uma das variáveis e aquelas que não possuíam observações foram excluídas.

A variável UF foi categorizada por região: norte, nordeste, centro-oeste, sul e sudeste, devido à pouca representatividade que apresentavam alguns estados.

Para a realização da modelagem preditiva no R, foram utilizados os pacotes Xgboost para a construção do modelo de árvore de classificação e a função glm para o modelo logístico com função de ligação “Logit”.

Quanto à variável dependente, ela foi mantida como 0 ou 1, sendo 1 quando o usuário tinha o desfecho e 0, caso contrário.

A base final foi dividida em treino e teste, sendo 75% das observações para treino e 25% para teste.

Após criada a base de treino, foi realizado o pré-processamento nos seguintes passos:

- 1- Criação da regra indicadora, onde determinamos que a variável “indicador” é a variável dependente e que as demais são variáveis preditoras;

- 2- Em seguida, as observações “na” das variáveis categóricas foram consideradas como desconhecidas para que elas não influenciassem o modelo;
- 3- As observações “na” das variáveis numéricas foram substituídas pela mediana das informações;
- 4- Foram excluídas do modelo as variáveis correlacionadas;
- 5- As variáveis categóricas foram transformadas em variáveis indicadoras (dummies).

Após o pré-processamento, os dados foram treinados utilizando-se das duas técnicas de classificação (logística e árvore de classificação) e, em seguida, foram preditos os resultados utilizando-se a base de teste.

4.7. Avaliação da qualidade dos modelos

Após a construção dos modelos logístico e da árvore de classificação, foi realizada a avaliação do poder preditivo de ambos os modelos. Para realizar a avaliação foram definidas as seguintes métricas: V verdadeiro, F falso, P positivo e N negativo.

O verdadeiro positivo e o verdadeiro negativo indicam que o valor previsto coincide com o valor observado correspondente a um sucesso ou a um fracasso, respectivamente. Já o falso positivo representa o Erro do Tipo I, quando um fracasso é classificado como um sucesso, e o falso negativo é caracterizado como o Erro do Tipo II, quando um sucesso é classificado como um fracasso.

Os indicadores utilizados para avaliar a qualidade do modelo, foram: Sensibilidade ($S = VP/(VP+FN)$), é a probabilidade do teste ter dado positivo (teve a doença) dado que o usuário teve a doença, quanto mais próximo de 1 melhor o modelo e quanto mais próximo de 0 pior; Especificidade ($E = VN/(VN+FP)$), é a probabilidade do teste ter dado negativo (não teve a doença) dado que o usuário não teve a doença, quanto mais próximo de 1 melhor o modelo e quanto mais próximo de 0 pior; Valor Preditivo Positivo ($VPP = VP/(VP+FP)$), é a probabilidade do usuário ter a doença dado que o teste deu positivo (teve a doença), quanto mais próximo de 1 melhor o modelo e quanto mais próximo de 0 pior; Valor Preditivo Negativo ($VPN = VN/(VN+FN)$), é a probabilidade do usuário não ter a doença dado que o teste tenha

dado negativo (não teve a doença), quanto mais próximo de 1 melhor o modelo e quanto mais próximo de 0 pior.

Baseado nos conceitos de sensibilidade e especificidade, a curva ROC é uma ferramenta gráfica utilizada para avaliar o poder de discriminação dos modelos. A área sob a curva ROC (AUC) é utilizada como resumo da curva ROC. É um valor que varia entre 0 e 1 e espera-se que os critérios de classificação tenham valor de AUC de, no mínimo, 0,50. Quanto mais próximo de 1, melhor é o poder preditivo do critério como um todo. Na literatura não existe um consenso sobre qual é o valor de AUC que determinaria uma boa capacidade de discriminação de um critério de classificação. Porém, Hosmer e colegas (2013) fornecem um critério de avaliação, o qual foi considerado neste trabalho (Quadro 4).

Quadro 4 – Critério de classificação da curva ROC (Hosmer et al., 2013)

AUC	Classificação
AUC = 0,5	Não possui poder de discriminação
$0,5 < \text{AUC} < 0,7$	Baixo poder de discriminação
$0,7 \leq \text{AUC} < 0,8$	Aceitável poder de discriminação
$0,8 \leq \text{AUC} < 0,9$	Excelente poder de discriminação
AUC $\geq 0,9$	Poder de discriminação acima do normal

Fonte: Hosmer et al., 2013

Por fim, foi definido o melhor ponto de corte através do indicador de eficiência. O indicador é obtido através da média aritmética entre a sensibilidade e a especificidade, sendo que o melhor ponto de corte será aquele que tiver a maior eficiência.

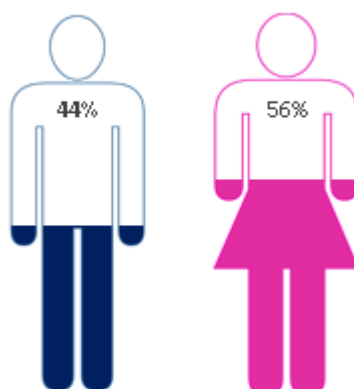
5. RESULTADOS E DISCUSSÃO

Neste capítulo, são descritas as características da população total analisada, dos atendimentos e procedimentos realizados e dos CIDs registrados em laudo segregado por ambulatorial e hospitalar. Por fim, é apresentado o resultado do processo de escolha do desfecho a ser predito. Em seguida, as características da população que apresentou o desfecho escolhido e o ajuste dos modelos preditivos comparados, além das análises de qualidade desses ajustes.

5.1. População estudada

Por meio da análise dos dados, foi possível observar que as mulheres representam 58,92% dos registros no período, enquanto os homens são responsáveis por 41,08% dos registros, conforme mostra o Gráfico 6.

Gráfico 6 – Proporção de utilização do SUS por sexo, Brasil, total de 2008 a 2018



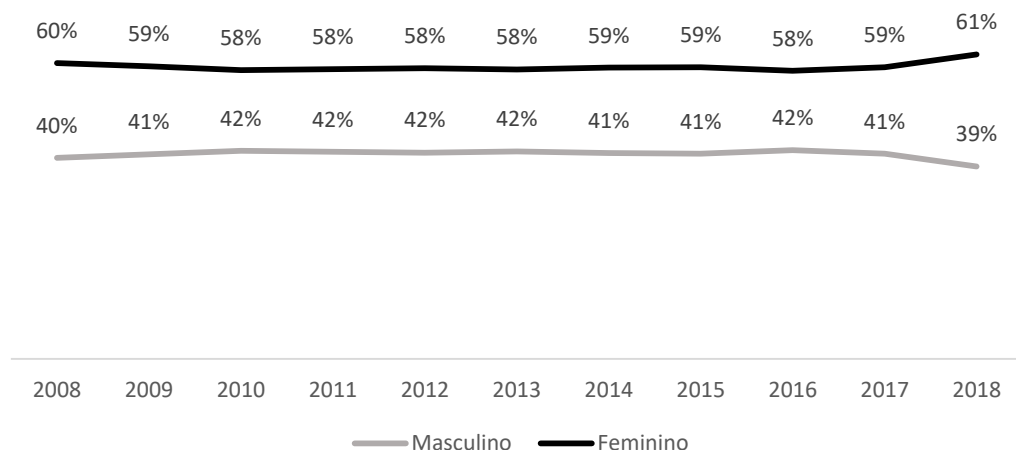
Fonte: Sistema de Informação DATASUS

Quando avaliamos a distribuição de frequências dos registros segundo sexo durante todo o recorte temporal, observamos que não existem diferenças muito grandes de um ano para outro, e é possível identificar que as utilizações femininas ficam sempre no patamar de 60% e, para os homens, em 40%, conforme aponta o Gráfico 7.

A diferença de registros entre homens e mulheres é notório, contudo não apresenta resultados destoantes da literatura e nem significativamente distantes um dos outros. Já é de conhecimento que mulheres utilizam mais do que os homens dos serviços de saúde, o que acaba provocando padrões de consumo diferentes entre os sexos (LEVORATO, 2013).

Desta forma, dentro dos critérios para a escolha do desfecho, foi considerada uma doença que atingisse tanto homens quanto mulheres, uma vez que uma enfermidade que atingisse apenas um dos sexos ensejaria uma perda significativa de observações dentro da base de dados.

Gráfico 7 – Proporção de utilização do SUS por sexo, Brasil, 2008 a 2018

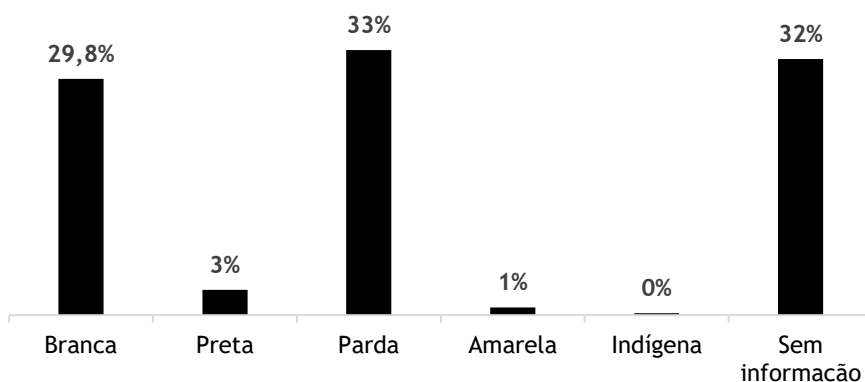


Fonte: Sistema de Informações – DATASUS

Ainda em relação às características dos usuários, o Gráfico 8 mostra a proporção de utilizações segundo a raça/cor autodeclarada. A maior parte da população estudada se autodeclara branca ou parda, sendo muito baixas as proporções de pessoas pretas, amarelas e indígenas.

Destaca-se a quantidade de informações faltantes na variável. Além disso, por se tratar de uma variável autodeclarada, as informações extraídas podem estar enviesadas, ou seja, o usuário se declara de uma raça/cor que, em termos de heteroidentificação, pode não condizer com sua realidade.

Gráfico 8 – Proporção de utilização por Raça/Cor, Brasil, 2008 a 2018

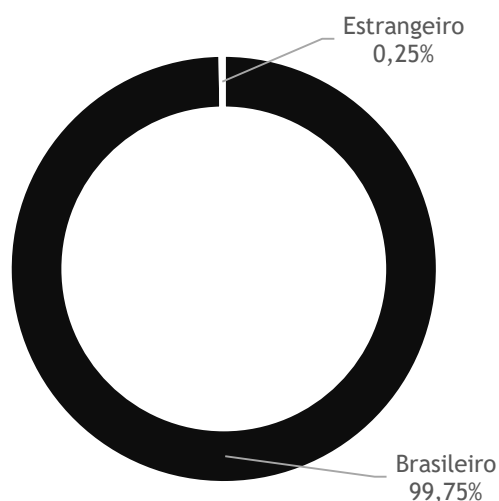


Fonte: Sistema de Informações – DATASUS

Pelo fato de a variável possuir um viés de auto identificação, ela não foi levada em consideração na modelagem preditiva e, por consequência, não foi preponderante na escolha do desfecho.

O Gráfico 9 apresenta as proporções de utilização do SUS, entre 2008 e 2018, por nacionalidade. A imensa maioria dos usuários eram de nacionalidade brasileira e as pessoas estrangeiras não representaram uma população significativamente importante para entrar nos modelos preditivos. Além disso, os modelos estão sendo criados para trabalharem nos contextos dos planos privados de saúde, nos quais a nacionalidade não é uma variável importante.

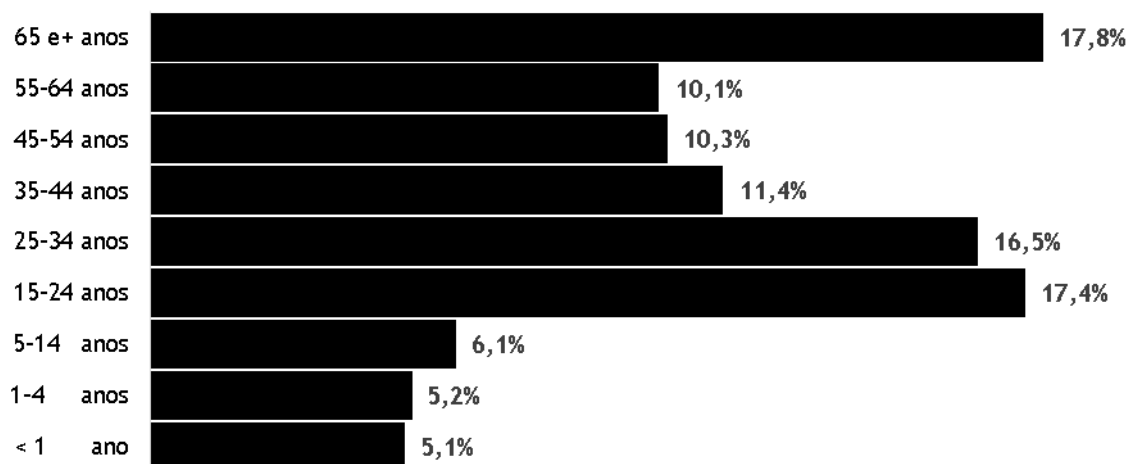
Gráfico 9 – Proporção de utilização dos SUS por Nacionalidade, Brasil, 2008 a 2018



Fonte: Sistema de Informações – DATASUS

Em termos gerais, os usuários do SUS apresentam um perfil etário mais envelhecido, em que a utilização da população de 0 a 24 anos representa 33,8% contra 38,2% da população com idade acima de 45 anos. Conforme evidencia o Gráfico 10.

Gráfico 10 – Distribuição etária dos usuários do SUS no Brasil de 2008 a 2018



Fonte: Sistema de Informações – DATASUS

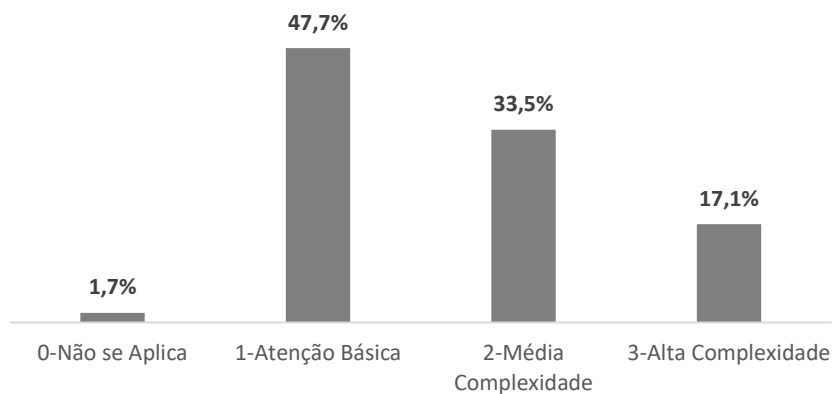
No processo de escolha do desfecho, levou-se em conta uma doença que fosse prevalente em todas as faixas etárias, uma vez que o Brasil vivencia em cada região distintos padrões de utilização por idade.

5.2. Perfil do atendimento prestado

Além das características que os bancos apresentam a respeito dos usuários do SUS, as variáveis das bases ambulatoriais e hospitalares que tratam sobre o tipo de atendimento que foi prestado são essenciais na análise, a fim de identificar padrões que possam auxiliar na escolha do desfecho e na modelagem preditiva.

Na base ambulatorial, uma das informações de extrema relevância é a complexidade do procedimento realizado. O Gráfico 11 apresenta a distribuição de frequência dos procedimentos realizados segundo o grau de complexidade. É possível notar que boa parcela dos procedimentos (48%) é voltado para a atenção básica.

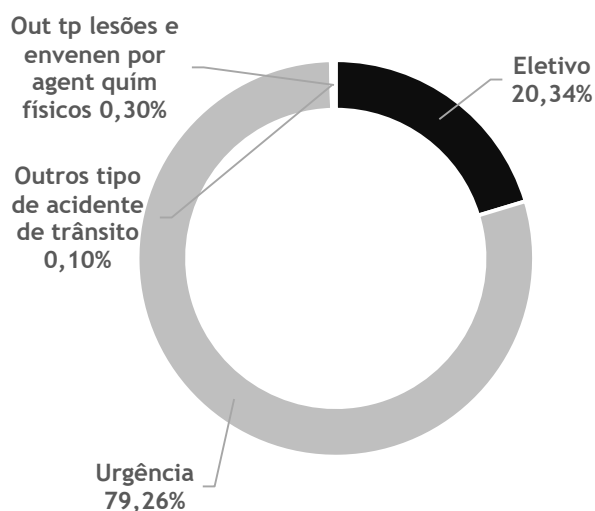
Gráfico 11 – Complexidade dos procedimentos ambulatoriais, Brasil, 2008 a 2018



Fonte: Sistema de Informações – DATASUS

O Gráfico 12 apresenta a distribuição de frequência das internações segundo seu caráter. A maior parte das internações tem caráter de urgência (79,3%). O restante das internações se dividem em eletivas (20,3%), por outros tipos de lesões e envenenamento por agentes químicos e físicos (0,3%) e outros tipos de acidentes de trânsito (0,1%).

Gráfico 12 – Caráter da Internação, Brasil, 2008 a 2018



Fonte: Sistema de Informações – DATASUS

5.3. Perfil dos procedimentos e diagnósticos dos usuários

Das bases de ocorrências ambulatoriais, os procedimentos relacionados à atenção básica, como visita domiciliar por profissional de nível médio, consulta médica em atenção primária e aferição de pressão arterial são os que possuem maior utilização, em torno de 2 bilhões em todo o recorte temporal de 2008 até 2018.

Em relação ao valor dos procedimentos ambulatoriais, o que possuiu maior valor bruto foi a Hemodiálise (máximo de 3 sessões por semana), que teve um custo de 17 bilhões em todo o recorte temporal e representou 13% dos custos ambulatoriais totais. Em seguida, os maiores custos são em consulta médica em atenção especializada e atendimento de urgência em atenção especializada.

Quando são avaliados os diagnósticos dos laudos ambulatoriais, ou seja, os CIDs, identificamos que o CID N18, referente a Insuficiência Renal Crônica, é o mais frequente dentro da população estudada, o que faz convergência com a informação de procedimentos mais caros, uma vez que a Hemodiálise é um dos tratamentos realizados para usuários com Insuficiência Renal.

Na análise dos CID com maior custo para tratamento, o N18 apareceu novamente em primeiro lugar no *ranking*, em seguida os CIDs C61 e C50.9, que são referentes a neoplasias malignas da próstata e neoplasia maligna da mama, respectivamente. Esses três CIDs representaram 13 bilhões de reais em gastos para tratamento.

Já nas bases de informações hospitalares, as internações mais frequentes foram relacionadas a tratamento de pneumonias ou influenza (gripe), parto normal e parto cesariano.

Em relação aos custos totais, os maiores gastos foram relacionados a tratamento com pneumonia ou influenza (gripe), tratamento com doenças bacterianas, parto normal e cesárea, o que representou cerca de 1,5 bilhões de reais para tratamento.

Avaliando os diagnósticos registrados nas internações pelo SUS, as Tabelas 2 e 3 apresentam as dez maiores quantidades de utilização e custo de tratamento em todo o recorte temporal de 2008 até 2018.

Tabela 2 – Categorias de CID com as 10 maiores quantidades de utilizações do SUS, de 2008 a 2018

CID 10 - CATEGORIA	QUANTIDADE
O80 Parto único espontâneo	536.553
J18 Pneumonia p/microrganismos NE	394.175
O82 Parto único p/cesariana	240.945
I50 Insuficiência cardíaca	203.417
J44 Outras doenças pulmonares obstrutivas crônicas	181.742
K80 Colelitíase	140.513
I20 Angina pectoris	135.980
J15 Pneumonia bacteriana NCOP	105.936
K40 Hérnia inguinal	99.425
I64 Acidente vascular cerebral NE como hemorragia isquêmico	92.736

Fonte: Sistema de Informações – DATASUS

Tabela 3 – Categorias de CID com os 10 maiores custos de tratamento do SUS, de 2008 a 2018

CID 10 - CATEGORIA	Custo
I20 Angina pectoris	R\$ 522.627.646
J18 Pneumonia p/microrganismos NE	R\$ 356.759.680
I21 Infarto agudo do miocárdio	R\$ 308.205.269
O80 Parto único espontâneo	R\$ 299.755.926
I50 Insuficiência cardíaca	R\$ 255.647.946
N18 Insuficiência renal crônica	R\$ 254.645.159
A41 Outras septicemias	R\$ 253.564.021
J96 Insuficiência respiratória NCOP	R\$ 203.105.953
P07 Transtornos relacionados com a gestação de curta duração e peso baixo ao nascer não classificados em outra parte	R\$ 194.608.812
O82 Parto único p/cesariana	R\$ 179.143.681

Fonte: Sistema de Informações – DATASUS

Destaca-se a prevalência de doenças crônicas, entre elas, o CID N18, que estava presente nos dados ambulatoriais, com um custo em internação de cerca de 254 milhões de reais para tratamento.

5.4. Escolha do desfecho a ser estudado

Baseado no perfil da população estudada, foi escolhido como desfecho para este estudo Insuficiência Renal Crônica (CID – N18), uma vez que esta doença atinge tanto homens como mulheres e está presente em todas as idades, apesar de ser mais frequente na população adulta.

Além disso, é uma doença crônica, que enseja tratamentos sofisticados e de longa duração, e está entre as doenças mais frequentes nos dados ambulatoriais, além de estar no *ranking* das dez doenças com maior custo de tratamento das bases ambulatoriais e hospitalares.

A importância de se prever antecipadamente a doença renal crônica, está associada ao fato de ela já ser considerada um dos importantes problemas médicos e de saúde pública no Brasil. Dados da Sociedade Brasileira de Nefrologia indicam que o número de doentes renais no Brasil dobrou na última década e estima que 10 milhões de brasileiros sofrem com alguma disfunção renal, além disso, entre 90 a 100 mil pessoas passam por diálise semanalmente no país (JUNIOR, 2004).

Segundo Guedes e colegas (2012), a Insuficiência Renal Crônica (IRC) é uma doença progressiva, debilitante, que causa incapacidade e apresenta alta mortalidade, além de que sua incidência e prevalência têm aumentado na população mundial.

A IRC é uma doença na qual os rins perdem a sua capacidade de efetuar suas funções básicas. Ela pode ser aguda quando acontece de forma súbita e rápida da perda das funções renais ou pode ser crônica, quando a perda acontece de forma lenta, irreversível e progressiva (PINHEIRO, 2011).

Para Junior (2004), a detecção precoce da doença renal e tratamentos terapêuticos para retardada sua progressão pode reduzir os sofrimentos dos usuários e os custos financeiros associados ao tratamento da doença. As duas principais causas da IRC é a hipertensão arterial e a diabetes mellitus. Além disso, as disfunções renais apresentam quase sempre uma evolução progressiva e assintomática, o que dificulta o diagnóstico preciso e precoce dos médicos.

Por fim, um dos pontos que foram fundamentais na escolha da IRC como desfecho a ser predito está na ocorrência de disfunção renal como uma das sequelas em indivíduos que tiveram Covid-19. Isso porque as células renais têm receptores

para o coronavírus semelhantes àqueles que existem nas células do pulmão, possibilitando então a infecção desses usuários (Pecly et al., 2021).

5.5. Perfil da população com Insuficiência Renal Crônica (IRC)

Após selecionado a IRC como desfecho do estudo, foi feito um primeiro processamento dos dados brutos a fim de avaliar quais eram os procedimentos e diagnósticos dos usuários do SUS antes de adoecerem.

A base final é composta pelas observações de 1,8 milhões de usuários do SUS, sendo que destes 198.748 adoeceram por IRC (11%) e 1,6 milhões não (89%).

Da população que foi diagnosticada com o CID N18 cerca de 45% eram mulheres e 55% eram homens. Em relação ao estado em que viviam, a maior parcela dos usuários estava concentrada em São Paulo (31%), conforme aponta Tabela 4.

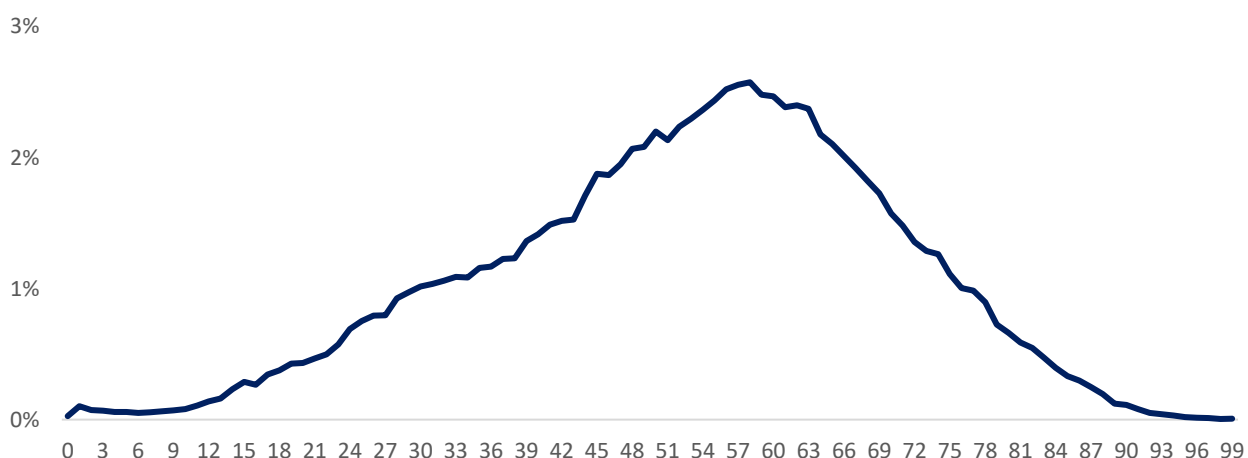
Tabela 4 – Proporção de usuários do SUS diagnosticados com IRC por estado, de 2008 a 2018

ESTADO	FREQUÊNCIA	FREQ POP. COM IRC	ESTADO	FREQUÊNCIA	FREQ POP. COM IRC
AC	0%	0,06%	PB	1%	0,02%
AL	1%	0,06%	PE	4%	0,08%
AM	1%	0,04%	PI	1%	0,08%
AP	0%	0,01%	PR	6%	0,29%
BA	5%	0,06%	RJ	6%	0,07%
CE	3%	0,08%	RN	1%	0,07%
DF	1%	0,07%	RO	1%	0,13%
ES	2%	0,08%	RR	0%	0,09%
GO	2%	0,07%	RS	8%	1,2%
MA	1%	0,04%	SC	4%	1,1%
MG	15%	4,1%	SE	1%	0,06%
MS	1%	0,09%	SP	31%	3,91%
MT	1%	0,05%	TO	0%	0,05%
PA	1%	0,03%	-	-	

Fonte: Sistema de Informações – DATASUS

A idade mediana dos usuários no momento em que são diagnosticados com insuficiência renal foi de 52 anos, e apesar de atingir todas as idades, ela apresenta ser uma doença que tem maior prevalência na população adulta, como pode ser observado no Gráfico 13.

Gráfico 13 – Distribuição de frequências dos usuários do SUS diagnosticados com IRC de 2008 a 2018



Fonte: Sistema de Informações – DATASUS

Em relação aos procedimentos realizados pelos usuários com IRC, destaca-se o acompanhamento de usuário pós-transplante de rim, fígado, coração, pulmão, células-tronco, hematopoiéticas e/ou pâncreas, que representa cerca de 40% do total de registros. Além disso, foram observados outros registros, que são os apresentados na Tabela 5.

Tabela 5 – Procedimentos realizados por usuários diagnosticados com IRC, de 2008 a 2018

Procedimento	Descrição	(%)
211020010	CATETERISMO CARDIACO	2%
211070319	SELECAO E VERIFICACAO DE BENEFICIO DO AASI	2%
301070032	ACOMPANHAMENTO DE USUÁRIO P/ ADAPTACAO DE APARELHO DE AMPLIF	3%
301080062	ACOMPANHAMENTO INTENSIVO DE USUÁRIO EM SAUDE MENTAL	3%
301080100	ACOMPANHAMENTO NAO INTENSIVO DE USUÁRIO EM SAUDE MENTAL	3%
301080127	ACOMPANHAMENTO SEMI-INTENSIVO DE USUÁRIOS EM SAUDE MENTAL	3%
301130019	AVALIACAO CLINICA E ELETRONICA DE DISPOSITIVO ELETRICO CARDIA	3%
309030129	LITOTRIPSIA EXTRACORPOREA (ONDA DE CHOQUE PARCIAL / COMPLETA	3%
405030045	FOTOCOAGULACAO A LASER	4%
405030193	PAN-FOTOCOAGULAÇÃO DE RETINA A LASER	4%
405050020	CAPSULOTOMIA A YAG LASER	4%
405050097	FACECTOMIA C/ IMPLANTE DE LENTE INTRA-OCULAR	4%
405050119	FACOEMULSIFICACAO C/ IMPLANTE DE LENTE INTRA-OCULAR RIGIDA	4%
405050372	FACOEMULSIFICACAO C/ IMPLANTE DE LENTE INTRA-OCULAR DOBRAVEL	4%
418010013	CONFECÇÃO DE FISTULA ARTERIO-VENOSA C/ ENXERTIA DE POLITETRAF	5%

Procedimento	Descrição	(%)
418010021	CONFECÇÃO DE FISTULA ARTERIO-VENOSA C/ ENXERTO AUTOLOGO	5%
418010080	IMPLANTE DE CATETER TIPO TENCKHOFF OU SIMILAR P/ DPA/DPAC	5%
418020019	INTERVENÇÃO EM FISTULA ARTERIO-VENOSA	5%
418020027	LIGADURA DE FISTULA ARTERIO-VENOSA	5%
506010023	ACOMPANHAMENTO DE USUÁRIO POS-TRANSPLANTE DE RIM FIGADO CORA	6%
506010040	ACOMPANHAMENTO DE USUÁRIOS NO PRÉ TRANSPLANTE DE ÓRGÃOS	6%
701030127	APARELHO DE AMPLIFICAÇÃO SONORA INDIVIDUAL (AASI) EXTERNO RET	8%
701030135	APARELHO DE AMPLIFICAÇÃO SONORA INDIVIDUAL (AASI) EXTERNO RET	8%

Fonte: Elaboração própria

Os procedimentos não apresentados aqui, mas que foram realizados pelos usuários, são aqueles que possuíam pequena variabilidade ou não apresentavam correlação com o desfecho estudado.

A Tabela 6 apresenta distribuição de frequências dos usuários com IRC segundo os diagnósticos apresentados antes do adoecimento pelo desfecho.

Tabela 6 – Diagnóstico por Capítulos do CID dos usuários do SUS antes de adoecerem com IRC, de 2008 a 2018

CAPITULO CID	DESCRIÇÃO	FREQUÊNCIA
F	Transtornos mentais e comportamentais	2,22%
H	Doenças do olho e anexos	16,32%
I	Doenças do aparelho circulatório	3,98%
N	Doenças do aparelho geniturinário	0,43%
Z	Fatores que exercem influência sobre o estado de saúde e o contato com serviços de saúde	76,73%

Fonte: Elaboração própria

Dentre os CIDs apresentados, é possível observar que cerca de 2,22% dos registros se referem a diagnósticos relacionados a adoecimentos mentais, como Esquizofrenia, Transtornos esquizotípico, Psicose não-orgânica, dentre outros adoecimentos do capítulo F - Transtornos mentais e comportamentais.

Além disso é possível identificar que 16,32% dos diagnósticos são do capítulo H, capítulo dos CIDs relacionados a Catarata, Retinopatia, Presbiacusia, além de outros cids do mesmo grupo, que se refere a doenças do olho e anexos. Uma das

possíveis causas para o aparecimento desses CIDs está nas idades em que a doença é mais prevalente, onde se pode associar o aparecimento da Insuficiência renal em conjunto a outras doenças relacionadas a visão que também ocorrem com maior frequência na população mais velha. Além de que a Diabetes que é uma doença que provoca lesões nos olhos, está intrinsicamente associada a IRC, sendo um dos sintomas de pacientes diagnosticados com insuficiência renal.

O aparecimento dos CIDs do capítulo I - Doenças do aparelho circulatório já era esperado, uma vez que este se refere as doenças cardíacas como a hipertensão que é característica dos usuários que desenvolvem quadros de IRC.

A mesma análise vale para as doenças do capítulo N - Doenças do aparelho geniturinário, que está associada aos sintomas que os usuários em geral apresentam antes do adoecimento por insuficiência renal.

Além disso, o capítulo Z - Fatores que influenciam o estado de saúde e o contato com os serviços de saúde é o que tem a maior representatividade (76,73%). O que pode ser explicado pela alta taxa de exames e observações por razões específicas que são realizados pelos usuários. Esses exames são frequentemente utilizados pelo corpo clínico para identificar a IRC.

Assim como nos procedimentos, alguns CIDs não foram apresentados aqui, apesar de aparecerem nos registros dos usuários. Estes não entraram no modelo por terem pouca variabilidade ou por não possuírem correlação com o desfecho estudado.

5.6. Medidas de qualidade dos modelos

Os dados de treino abrigavam 75% da base total, o que representa 1,3 milhões de informações, sendo que 11% dos usuários possuíam IRC (149.061) e 89% não (1.216.401). A base de teste corresponde aos 25% restante da base total (455.154), sendo que 11% dos usuários possuíam IRC (49.687) e 89% não (405.467).

A tabela 7 mostra as medidas de qualidade para os dois modelos sob os diferentes pontos de corte usados para definir o que é um positivo a partir da probabilidade predita. O Apêndice A apresenta as matrizes com os resultados dos Verdadeiros Positivos (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) para o modelo logístico e árvore de classificação.

Tabela 7 – Medidas de qualidade dos modelos de regressão logística e de árvore de decisão sob os diferentes pontos de corte

Ponto de corte para a probabilidade predita	Árvore de Classificação				Regressão Logística			
	Sensibilidade (S)	Especificidade (E)	Valor Preditivo Positivo (VPP)	Valor Preditivo Negativo (VPN)	Sensibilidade (S)	Especificidade (E)	Valor Preditivo Positivo (VPP)	Valor Preditivo Negativo (VPN)
0,1	0,844	0,516	0,176	0,964	0,813	0,795	0,327	0,972
0,2	0,519	0,751	0,204	0,927	0,622	0,934	0,537	0,953
0,3	0,329	0,824	0,187	0,909	0,471	0,969	0,649	0,937
0,4	0,301	0,832	0,180	0,907	0,418	0,976	0,679	0,932
0,5	0,299	0,832	0,179	0,906	0,360	0,981	0,696	0,926
0,6	0,299	0,832	0,179	0,906	0,293	0,986	0,715	0,919
0,7	0,297	0,832	0,178	0,906	0,226	0,99	0,733	0,913
0,8	0,297	0,832	0,178	0,906	0,130	0,995	0,748	0,903
0,9	0,296	0,833	0,178	0,906	0,055	0,998	0,807	0,896

Fonte: Elaboração Própria

Como resultado, nota-se que, o modelo de árvore de classificação prevê um número maior de usuários que deram positivo dentre os usuários que de fato tiveram a doença, ou seja, possui uma melhor sensibilidade do que o logístico, para a maior parte dos pontos de corte.

A melhor técnica levando-se em conta a especificidade é a logístico, dado que ela tem maior poder preditivo para identificar os usuários que são de fato saudáveis (o teste deu negativo) dentre aqueles que não tiveram a doença.

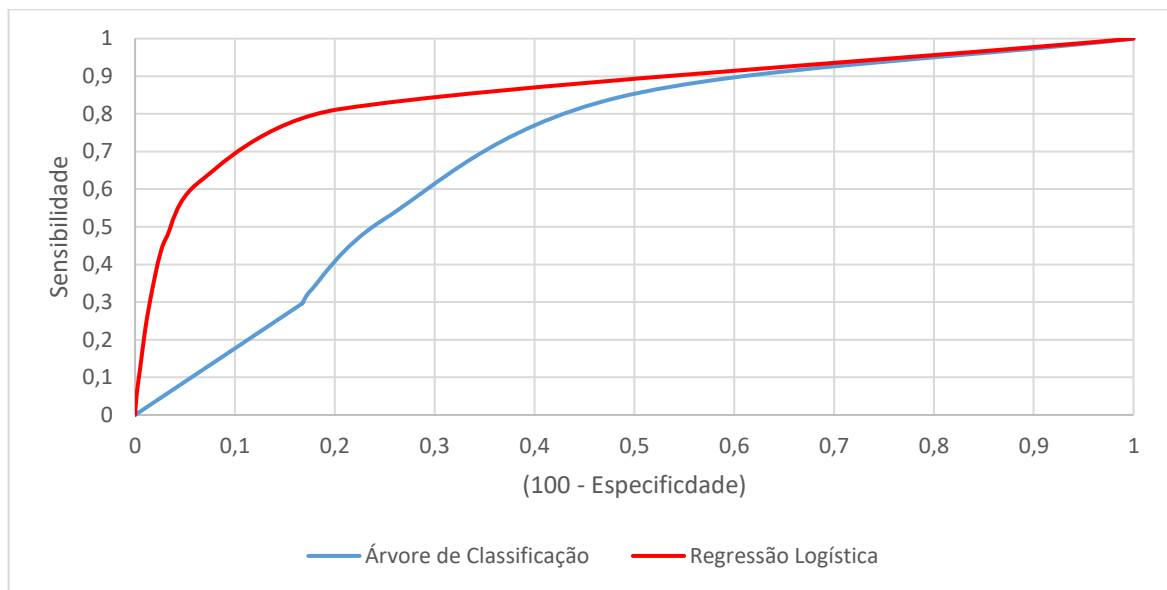
Em relação aos resultados do Valor Preditivo Positivo o modelo logístico se caracteriza como a melhor técnica, com resultados bem maiores que os apresentados pelo modelo de árvore de classificação. Este resulta exemplifica que o logístico tem maior capacidade de detectar os usuários que de fato tiveram a doença dentre aqueles que a predição apontou que adoeceram por IRC (teste deu positivo).

Quando avaliamos os modelos levando-se em conta o Valor Preditivo Negativo, o modelo logístico é melhor em identificar os usuários que de fato não tiveram a doença dentre aqueles que a predição apontou que eram saudáveis (teste deu negativo).

Para este estudo o ponto de corte escolhido para os dois modelos foi o de 0,1 em que a eficiência do modelo de regressão logística foi de 0,804 e para o de árvore de classificação foi de 0,680.

Baseando-se nos resultados encontrados de sensibilidade e especificidade foram geradas as curvas ROC que são apresentadas no gráfico 14 para avaliar a qualidade das duas técnicas em distinguir entre as duas classificações: adoeceu ou não adoeceu por IRC.

Gráfico 14 – Curva ROC dos modelos de regressão logística e de árvore de classificação



Fonte: Elaboração Própria

O modelo de regressão logística teve uma performance em termos da área abaixo a curva ROC de 79,49%, que de acordo com Hosmer o modelo possui aceitável poder de discriminação.

O modelo de árvore de classificação teve uma performance em termos da área abaixo a curva ROC de 51,59%, que de acordo com Hosmer o modelo possui baixo poder de discriminação.

Os resultados da AUC exemplificam a superioridade do modelo logístico em classificar os usuários em relação ao de árvore de classificação.

6. CONCLUSÃO

A saúde no Brasil vem enfrentando diversas problemáticas relacionadas ao aumento dos custos médico hospitalares, chamando atenção do mercado para procurar modelos para evitar a alta das sinistralidades ou insolvência das operadoras.

Um caminho para se pensar em saúde de forma preventiva é a adoção de modelos preditivos, que têm por objetivo trazer um olhar individualizado para os beneficiários, mudando o foco da doença para a saúde dos usuários, impactando na redução dos custos, mas sobretudo no aumento da qualidade de vida dos indivíduos.

Em termos do impacto da adoção de modelos preditivos nas operadoras de planos de saúde, se destaca principalmente a redução nos custos com os beneficiários que possuem doenças que têm tratamento mais onerosos, uma vez que, identificando os beneficiários que possuem maior propensão a adoecerem por doenças que ensejam altos custos com tratamento em ambulatório e hospital, as operadoras vão articular políticas internas para impedir ou retardar o avanço da doença.

Através dos resultados apresentados conseguimos destacar a superioridade do modelo logístico em relação ao modelo por árvore de classificação. O modelo logístico possui maior área abaixo da curva, destacando-se com um aceitável poder de discriminação em relação ao poder de discriminação da árvore de classificação.

Em relação aos indicadores de performance dos modelos, o logístico apresentou maior especificidade para todos os pontos de corte, destacando-se na detecção dos usuários que de fato não tiveram a doença.

O modelo de árvore de classificação possui sensibilidade mais alta em relação ao logístico na maior parte dos pontos de corte, porém, os dois modelos apresentam sensibilidade baixa a partir do ponto de corte 0,2. Esse resultado aponta para o baixo poder em identificar os usuários que tiveram a doença (teste deu positivo) dentre aqueles que de fato tiveram a doença.

Os dois modelos tiveram resultados satisfatórios do Valor Preditivo Negativo para todos os pontos de corte, destacando-se com alto poder para identificar os usuários que não tiveram a doença dentre aqueles que a predição deu negativo.

O modelo logístico teve melhor performance em relação ao Poder Preditivo Positivo, destacando-se na identificação dos usuários que tiveram a doença dentre aqueles que o teste deu positivo.

Dentre os pontos de cortes utilizados no estudo, o corte escolhido para a modelagem preditiva de IRC foi baseado na eficiência, e aquele que apresentou melhor resultado foi o de 0,1 para ambos os modelos.

Os dados utilizados no estudo eram desbalanceados, em que cerca de 11% eram referentes a usuários que adoeceram por IRC e os 89% restantes de usuários que adoeceram por outras causas. O desbalanceamento provoca um efeito na predição em que os modelos conhecem com uma maior precisão os usuários que não adoeceram pelo desfecho, do que os que de fato adoeceram.

Para os próximos trabalhos, espera-se aumentar o número de doenças a serem preditas, a fim de que cada usuário, seja do SUS como do sistema privado, possa ter um ranqueamento das doenças que possui maior risco de adoecimento. Além da realização de outras modelagens com a utilização de dados balanceados a fim de verificar se existe melhora no poder preditivo dos modelos, além da utilização de outras técnicas de predição.

7. REFERENCIA BIBLIOGRÁFICA

ALBUQUERQUE, Ceres. **A situação atual do mercado da saúde suplementar no Brasil e apontamentos para o futuro.** 2007. Disponível: <<https://www.scielo.br/j/csc/a/jXwhKzH5MtFjLS4h7WdMy8m/?lang=pt>>. Acesso em: 06/01/2022

ANDRADE, Eli et al.,. **Gênese de uma política pública de ações de alto custo e complexidade: As Terapias Renais Substitutivas no Brasil.** 2006. Disponível: <https://www.nescon.medicina.ufmg.br/biblioteca/registro/Genese_de_uma_politica_publica_de_acoes_de_alto_custo_e_complexidade__As_Terapias_Renais_Substitutivas_no_Brasil_/357>. Acesso em: 20/01/2022

ARAÚJO, Ângelo et al., **Análise de tendência da sinistralidade e impacto na diminuição do número de operadoras de saúde suplementar no Brasil.**2016. Disponível: <<https://www.scielo.br/j/csc/a/zmVZr3B95wgTph7zcBJMSCc/?format=pdf&lang=pt>>. Acesso em: 12/01/2022

CARVALHO, Eurípedes et al.,. **A regulamentação do setor de saúde suplementar no Brasil: a reconstrução de uma história de disputas.** 2007. Disponível: <<https://www.scielo.br/j/csp/a/G639bDbmszZqRwYJqFh9ZRb/abstract/?lang=pt>>. Acesso em: 10/01/2022

COTA, Isamara et al., . **ANÁLISE DA EVOLUÇÃO DOS CUSTOS ASSISTENCIAIS DAS OPERADORAS DE PLANOS DE SAÚDE CONSIDERANDO FATOR MODERADOR.** 2019. Disponível: <<https://anaiscbc.emnuvens.com.br/anais/article/view/4512>>. Acesso em: 15/01/2022

FERNANDES, Eder. **Investigação de Notificação de Evento Adverso Pós-Vacinação.** 2022. Disponível: <<https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/calendario-nacional-de-vacinacao/eventos-adversos-pos->

LIMA, Clóvis et al.,. **INFORMAÇÃO, ASSIMETRIA DE INFORMAÇÕES E REGULAÇÃO DO MERCADO DE SAÚDE SUPLEMENTAR.** 2006. Disponível: <<https://brapci.inf.br/index.php/res/download/96056>>. Acesso em: 15/01/2022

LIMA, Jean et al., . **Sinistralidade em contratos de plano de saúde médico hospitalar.** 2019. Disponível: <<https://periodicos.uninove.br/revistargss/article/view/14975>>. Acesso em: 14/01/2022

MACHADO, Pamila et al.,. **Insuficiência renal crônica e as causas múltiplas de morte: uma análise descritiva para o Brasil, 2000 a 2004.** 2014. Disponível: <https://www.scielo.br/scielo.php?pid=S1414462X2014000400372&script=sci_arttext>. Acesso em: 20/01/2022

MARTINS, Andriago. **Aplicação de Análise de Risco de Crédito com o uso das Técnicas de Regressão Logística e Árvores de Decisão.** 2020. Disponível: <<https://repositorio.ufmg.br/handle/1843/35524>>. Acesso em: 23/01/2022

MATOS, João. **PLANOS DE SAÚDE: UMA ANÁLISE DOS CUSTOS ASSISTENCIAIS E SEUS COMPONENTES.** 2009. Disponível: <<https://www.scielo.br/j/rae/a/DHWJphzZP3FmKdnGxkVnkjz/?format=pdf&lang=pt>>. Acesso em: 14/01/2022

MEIRA, Carlos et al., . **Análise da epidemia da ferrugem do cafeeiro com árvore de decisão.** 2008. Disponível: <<https://www.scielo.br/j/tpp/a/gwhNLwQB58hkwJSSWdJ7gbB/?lang=pt>>. Acesso em: 17/01/2022

MENDES, Antonio et al., . **Assistência pública de saúde no contexto da transição demográfica brasileira: exigências atuais e futuras.** 2012. Disponível: <<https://www.scielo.br/j/csp/a/YqW3NNYWrvmFWfVksfmLgpj/abstract/?lang=pt>>. Acesso em: 06/01/2022

MINUSSI, João et al., . **Um Modelo de Previsão de Solvência Utilizando Regressão Logística.** 2002. Disponível: <<https://www.scielo.br/j/rac/a/LBTBV5pcVXXg4Ks55LH83fr/?lang=pt#:~:text=Atrav%C3%A9s%20da%20aplica%C3%A7%C3%A3o%20da%20an%C3%A1lise,das%20empresas%20foram%20classificadas%20corretamente.>>. Acesso em: 16/01/2022

NASCIMENTO, Ginivaldo. **Síndrome Cardiorrenal Tipo 1 em Região de Baixo Desenvolvimento: Comparação entre os Critérios AKIN e KDIGO, Necessidade de Diálise e Mortalidade.** 2020. Disponível: <<https://www.scielo.br/j/abc/a/ChwGkyWyHFZqzhJfqvVJTmB/>>. Acesso em: 20/01/2022

OHTOSHI, Claudia. **Uma comparação de regressão logística, árvore de classificação e redes neurais: Analisando dados de crédito.** 2003. Disponível: <<https://www.teses.usp.br/teses/disponiveis/45/45133/tde-20210729132841/en.php>>. Acesso em: 17/01/2022

PIETROBON, Louise. **Saúde suplementar no Brasil: o papel da Agência Nacional de Saúde Suplementar na regulação do setor.** 2008. Disponível: <<https://www.scielo.br/j/physis/a/KFy6MMGRnjWVLNL7DKkXRKm/?lang=pt#:~:text=A%20Ag%C3%AAncia%20Nacional%20de%20Sa%C3%BAde,a%20assist%C3%AAncia%20suplementar%20%C3%A0%20sa%C3%BAde.>>. Acesso em: 11/01/2022

PIRES, Maria et al., . **Oferta e demanda por média complexidade/SUS: relação com atenção básica.** 2008. Disponível: <https://www.scielo.br/scielo.php?pid=S141381232010000700007&script=sci_arttext>. Acesso em: 20/01/2022

REZENDE, Paulo. **OS CONTRATOS DE PLANO DE SAÚDE E SEU EQUILÍBRIO ECONÔMICO-FINANCEIRO: MUTUALISMO, CÁLCULO ATUARIAL E O IMPACTO ECONÔMICO DAS DECISÕES JUDICIAIS.** 2011. Disponível: <<http://www3.mcampos.br:84/u/201503/paulorobertovogelderezendeoscontratosdeplanosaudeesequilibrioeconomicofinanceiro.pdf>>. Acesso em: 24/01/2022

SESTELO, José et al., . **Saúde suplementar no Brasil: abordagens sobre a articulação público/privada na assistência à saúde.** 2013. Disponível: <<https://www.scielo.br/pdf/csp/v29n5/04.pdf>>. Acesso em: 06/01/2022

TACONELI, Cesar. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia.** 2008. Disponível: <<http://www.leg.ufpr.br/lib/exe/fetch.php/projetos:modeltrees:tesecesarconeli.pdf>>. Acesso em: 17/01/2022

TEIXEIRA, Carmen. **Transição epidemiológica, modelo de atenção à saúde e previdência social no Brasil: problematizando tendências e opções políticas.** 2004. Disponível:<https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232004000400003>. Acesso em: 06/01/2022

THEMELSLE, **Didática Tech.Underfitting e Overfitting.** 2022. Disponível: <<https://didatica.tech/underfitting-e-overfitting/>>. Acesso em: 17/01/2022

APÊNDICE A

Descrição das variáveis

CÓDIGO	DESCRIÇÃO
COD_PAC	CÓDIGO DO USUÁRIO
INDICADOR	1 - COM IRC / 0- SEM IRC
SEXO	M/F
IDADE	NÚMERO CONTÍNUO
UF	27 ESTADOS BRASILEIROS MAIS O DF
QU2A	QUANTAS VEZES UTILIZOU 2A
QU2A_201	QUANTAS 201 REALIZOU NOS 2A
QU2A_206	QUANTAS 206 REALIZOU NOS 2A
QU2A_211	QUANTAS 211 REALIZOU NOS 2A
QU2A_301	QUANTOS 301 2A
QU2A_307	QUANTAS 307 2A
QU2A_309	QUANTAS 309 2A
QU2A_405	QUANTOS 405 REALIZOU 2A
QU2A_409	QUANTOS 409 REALIZOU 2A
QU2A_414	QUANTOS 414 REALIZOU 2A
QU2A_418	QUANTOS 418 REALIZOU 2A
QU2A_503	QUANTOS 503 REALIZOU 2A
QU2A_504	QUANTOS 504 REALIZOU 2A
QU2A_505	QUANTOS 505 REALIZOU 2A
QU2A_506	QUANTOS 506 REALIZOU 2A
QU2A_701	QUANTOS 701 REALIZOU 2A
QU2A_211020010	QUANTAS VEZES FEZ O PROCEDIMENTO (211020010) NOS ULTIMOS 2A
QU2A_211070319	QUANTAS VEZES FEZ O PROCEDIMENTO (211070319) NOS ULTIMOS 2A
QU2A_301070032	QUANTAS VEZES FEZ O PROCEDIMENTO (301070032) NOS ULTIMOS 2A
QU2A_301080062	QUANTAS VEZES FEZ O PROCEDIMENTO (301080062) NOS ULTIMOS 2A
QU2A_301080100	QUANTAS VEZES FEZ O PROCEDIMENTO (301080100) NOS ULTIMOS 2A
QU2A_301080127	QUANTAS VEZES FEZ O PROCEDIMENTO (301080127) NOS ULTIMOS 2A
QU2A_301130019	QUANTAS VEZES FEZ O PROCEDIMENTO (301130019) NOS ULTIMOS 2A
QU2A_309030129	QUANTAS VEZES FEZ O PROCEDIMENTO (309030129) NOS ULTIMOS 2A
QU2A_405030045	QUANTAS VEZES FEZ O PROCEDIMENTO (405030045) NOS ULTIMOS 2A
QU2A_405030193	QUANTAS VEZES FEZ O PROCEDIMENTO (405030193) NOS ULTIMOS 2A
QU2A_405050020	QUANTAS VEZES FEZ O PROCEDIMENTO (405050020) NOS ULTIMOS 2A
QU2A_405050097	QUANTAS VEZES FEZ O PROCEDIMENTO (405050097) NOS ULTIMOS 2A
QU2A_405050119	QUANTAS VEZES FEZ O PROCEDIMENTO (405050119) NOS ULTIMOS 2A
QU2A_405050372	QUANTAS VEZES FEZ O PROCEDIMENTO (405050372) NOS ULTIMOS 2A
QU2A_418010013	QUANTAS VEZES FEZ O PROCEDIMENTO (418010013) NOS ULTIMOS 2A
QU2A_418010021	QUANTAS VEZES FEZ O PROCEDIMENTO (418010021) NOS ULTIMOS 2A
QU2A_418010080	QUANTAS VEZES FEZ O PROCEDIMENTO (418010080) NOS ULTIMOS 2A
QU2A_418020019	QUANTAS VEZES FEZ O PROCEDIMENTO (418020019) NOS ULTIMOS 2A
QU2A_418020027	QUANTAS VEZES FEZ O PROCEDIMENTO (418020027) NOS ULTIMOS 2A
QU2A_506010023	QUANTAS VEZES FEZ O PROCEDIMENTO (506010023) NOS ULTIMOS 2A
QU2A_506010040	QUANTAS VEZES FEZ O PROCEDIMENTO (506010040) NOS ULTIMOS 2A
QU2A_701030127	QUANTAS VEZES FEZ O PROCEDIMENTO (701030127) NOS ULTIMOS 2A
QU2A_701030135	QUANTAS VEZES FEZ O PROCEDIMENTO (701030135) NOS ULTIMOS 2A
QU2A_F063	QUANTAS VEZES FEZ O PROCEDIMENTO (F063) NOS ULTIMOS 2A
QU2A_F069	QUANTAS VEZES FEZ O PROCEDIMENTO (F069) NOS ULTIMOS 2A
QU2A_F078	QUANTAS VEZES FEZ O PROCEDIMENTO (F078) NOS ULTIMOS 2A

CÓDIGO	DESCRIÇÃO
QU2A_F09	QUANTAS VEZES FEZ O PROCEDIMENTO (F09) NOS ULTIMOS 2A
QU2A_F102	QUANTAS VEZES FEZ O PROCEDIMENTO (F102) NOS ULTIMOS 2A
QU2A_F142	QUANTAS VEZES FEZ O PROCEDIMENTO (F142) NOS ULTIMOS 2A
QU2A_F192	QUANTAS VEZES FEZ O PROCEDIMENTO (F192) NOS ULTIMOS 2A
QU2A_F200	QUANTAS VEZES FEZ O PROCEDIMENTO (F200) NOS ULTIMOS 2A
QU2A_F201	QUANTAS VEZES FEZ O PROCEDIMENTO (F201) NOS ULTIMOS 2A
QU2A_F205	QUANTAS VEZES FEZ O PROCEDIMENTO (F205) NOS ULTIMOS 2A
QU2A_F21	QUANTAS VEZES FEZ O PROCEDIMENTO (F21) NOS ULTIMOS 2A
QU2A_F250	QUANTAS VEZES FEZ O PROCEDIMENTO (F250) NOS ULTIMOS 2A
QU2A_F29	QUANTAS VEZES FEZ O PROCEDIMENTO (F29) NOS ULTIMOS 2A
QU2A_F310	QUANTAS VEZES FEZ O PROCEDIMENTO (F310) NOS ULTIMOS 2A
QU2A_F316	QUANTAS VEZES FEZ O PROCEDIMENTO (F316) NOS ULTIMOS 2A
QU2A_F317	QUANTAS VEZES FEZ O PROCEDIMENTO (F317) NOS ULTIMOS 2A
QU2A_F319	QUANTAS VEZES FEZ O PROCEDIMENTO (F319) NOS ULTIMOS 2A
QU2A_F320	QUANTAS VEZES FEZ O PROCEDIMENTO (F320) NOS ULTIMOS 2A
QU2A_F321	QUANTAS VEZES FEZ O PROCEDIMENTO (F321) NOS ULTIMOS 2A
QU2A_F322	QUANTAS VEZES FEZ O PROCEDIMENTO (F322) NOS ULTIMOS 2A
QU2A_F323	QUANTAS VEZES FEZ O PROCEDIMENTO (F323) NOS ULTIMOS 2A
QU2A_F330	QUANTAS VEZES FEZ O PROCEDIMENTO (F330) NOS ULTIMOS 2A
QU2A_F331	QUANTAS VEZES FEZ O PROCEDIMENTO (F331) NOS ULTIMOS 2A
QU2A_F332	QUANTAS VEZES FEZ O PROCEDIMENTO (F332) NOS ULTIMOS 2A
QU2A_F412	QUANTAS VEZES FEZ O PROCEDIMENTO (F412) NOS ULTIMOS 2A
QU2A_F432	QUANTAS VEZES FEZ O PROCEDIMENTO (F432) NOS ULTIMOS 2A
QU2A_F600	QUANTAS VEZES FEZ O PROCEDIMENTO (F600) NOS ULTIMOS 2A
QU2A_H250	QUANTAS VEZES FEZ O PROCEDIMENTO (H250) NOS ULTIMOS 2A
QU2A_H251	QUANTAS VEZES FEZ O PROCEDIMENTO (H251) NOS ULTIMOS 2A
QU2A_H258	QUANTAS VEZES FEZ O PROCEDIMENTO (H258) NOS ULTIMOS 2A
QU2A_H259	QUANTAS VEZES FEZ O PROCEDIMENTO (H259) NOS ULTIMOS 2A
QU2A_H260	QUANTAS VEZES FEZ O PROCEDIMENTO (H260) NOS ULTIMOS 2A
QU2A_H264	QUANTAS VEZES FEZ O PROCEDIMENTO (H264) NOS ULTIMOS 2A
QU2A_H268	QUANTAS VEZES FEZ O PROCEDIMENTO (H268) NOS ULTIMOS 2A
QU2A_H269	QUANTAS VEZES FEZ O PROCEDIMENTO (H269) NOS ULTIMOS 2A
QU2A_H330	QUANTAS VEZES FEZ O PROCEDIMENTO (H330) NOS ULTIMOS 2A
QU2A_H340	QUANTAS VEZES FEZ O PROCEDIMENTO (H340) NOS ULTIMOS 2A
QU2A_H360	QUANTAS VEZES FEZ O PROCEDIMENTO (H360) NOS ULTIMOS 2A
QU2A_H368	QUANTAS VEZES FEZ O PROCEDIMENTO (H368) NOS ULTIMOS 2A
QU2A_H402	QUANTAS VEZES FEZ O PROCEDIMENTO (H402) NOS ULTIMOS 2A
QU2A_H900	QUANTAS VEZES FEZ O PROCEDIMENTO (H900) NOS ULTIMOS 2A
QU2A_H903	QUANTAS VEZES FEZ O PROCEDIMENTO (H903) NOS ULTIMOS 2A
QU2A_H904	QUANTAS VEZES FEZ O PROCEDIMENTO (H904) NOS ULTIMOS 2A
QU2A_H905	QUANTAS VEZES FEZ O PROCEDIMENTO (H905) NOS ULTIMOS 2A
QU2A_H906	QUANTAS VEZES FEZ O PROCEDIMENTO (H906) NOS ULTIMOS 2A
QU2A_H908	QUANTAS VEZES FEZ O PROCEDIMENTO (H908) NOS ULTIMOS 2A
QU2A_H911	QUANTAS VEZES FEZ O PROCEDIMENTO (H911) NOS ULTIMOS 2A
QU2A_H918	QUANTAS VEZES FEZ O PROCEDIMENTO (H918) NOS ULTIMOS 2A
QU2A_H919	QUANTAS VEZES FEZ O PROCEDIMENTO (H919) NOS ULTIMOS 2A
QU2A_H932	QUANTAS VEZES FEZ O PROCEDIMENTO (H932) NOS ULTIMOS 2A
QU2A_I050	QUANTAS VEZES FEZ O PROCEDIMENTO (I050) NOS ULTIMOS 2A
QU2A_I120	QUANTAS VEZES FEZ O PROCEDIMENTO (I120) NOS ULTIMOS 2A
QU2A_I200	QUANTAS VEZES FEZ O PROCEDIMENTO (I200) NOS ULTIMOS 2A
QU2A_I201	QUANTAS VEZES FEZ O PROCEDIMENTO (I201) NOS ULTIMOS 2A

CÓDIGO	DESCRIÇÃO
QU2A_I208	QUANTAS VEZES FEZ O PROCEDIMENTO (I208) NOS ULTIMOS 2A
QU2A_I209	QUANTAS VEZES FEZ O PROCEDIMENTO (I209) NOS ULTIMOS 2A
QU2A_I210	QUANTAS VEZES FEZ O PROCEDIMENTO (I210) NOS ULTIMOS 2A
QU2A_I219	QUANTAS VEZES FEZ O PROCEDIMENTO (I219) NOS ULTIMOS 2A
QU2A_I220	QUANTAS VEZES FEZ O PROCEDIMENTO (I220) NOS ULTIMOS 2A
QU2A_I248	QUANTAS VEZES FEZ O PROCEDIMENTO (I248) NOS ULTIMOS 2A
QU2A_I249	QUANTAS VEZES FEZ O PROCEDIMENTO (I249) NOS ULTIMOS 2A
QU2A_I250	QUANTAS VEZES FEZ O PROCEDIMENTO (I250) NOS ULTIMOS 2A
QU2A_I251	QUANTAS VEZES FEZ O PROCEDIMENTO (I251) NOS ULTIMOS 2A
QU2A_I255	QUANTAS VEZES FEZ O PROCEDIMENTO (I255) NOS ULTIMOS 2A
QU2A_I258	QUANTAS VEZES FEZ O PROCEDIMENTO (I258) NOS ULTIMOS 2A
QU2A_I259	QUANTAS VEZES FEZ O PROCEDIMENTO (I259) NOS ULTIMOS 2A
QU2A_I442	QUANTAS VEZES FEZ O PROCEDIMENTO (I442) NOS ULTIMOS 2A
QU2A_I443	QUANTAS VEZES FEZ O PROCEDIMENTO (I443) NOS ULTIMOS 2A
QU2A_I498	QUANTAS VEZES FEZ O PROCEDIMENTO (I498) NOS ULTIMOS 2A
QU2A_I499	QUANTAS VEZES FEZ O PROCEDIMENTO (I499) NOS ULTIMOS 2A
QU2A_N200	QUANTAS VEZES FEZ O PROCEDIMENTO (N200) NOS ULTIMOS 2A
QU2A_N201	QUANTAS VEZES FEZ O PROCEDIMENTO (N201) NOS ULTIMOS 2A
QU2A_N47	QUANTAS VEZES FEZ O PROCEDIMENTO (N47) NOS ULTIMOS 2A
QU2A_Z048	QUANTAS VEZES FEZ O PROCEDIMENTO (Z048) NOS ULTIMOS 2A
QU2A_Z524	QUANTAS VEZES FEZ O PROCEDIMENTO (Z524) NOS ULTIMOS 2A
QU2A_Z525	QUANTAS VEZES FEZ O PROCEDIMENTO (Z525) NOS ULTIMOS 2A
QU2A_Z940	QUANTAS VEZES FEZ O PROCEDIMENTO (Z940) NOS ULTIMOS 2A
QU2A_Z941	QUANTAS VEZES FEZ O PROCEDIMENTO (Z941) NOS ULTIMOS 2A
QU2A_Z942	QUANTAS VEZES FEZ O PROCEDIMENTO (Z942) NOS ULTIMOS 2A
QU2A_Z944	QUANTAS VEZES FEZ O PROCEDIMENTO (Z944) NOS ULTIMOS 2A
QU2A_Z947	QUANTAS VEZES FEZ O PROCEDIMENTO (Z947) NOS ULTIMOS 2A
QU2A_Z948	QUANTAS VEZES FEZ O PROCEDIMENTO (Z948) NOS ULTIMOS 2A
QU2A_E11	QUANTAS VEZES FEZ O PROCEDIMENTO (E11) NOS ULTIMOS 2A
QU2A_E115	QUANTAS VEZES FEZ O PROCEDIMENTO (E115) NOS ULTIMOS 2A
QU2A_E116	QUANTAS VEZES FEZ O PROCEDIMENTO (E116) NOS ULTIMOS 2A
QU2A_E112	QUANTAS VEZES FEZ O PROCEDIMENTO (E112) NOS ULTIMOS 2A
QU2A_I15	QUANTAS VEZES FEZ O PROCEDIMENTO (I15) NOS ULTIMOS 2A
QU2A_I151	QUANTAS VEZES FEZ O PROCEDIMENTO (I151) NOS ULTIMOS 2A
QU2A_I152	QUANTAS VEZES FEZ O PROCEDIMENTO (I152) NOS ULTIMOS 2A
QU2A_I159	QUANTAS VEZES FEZ O PROCEDIMENTO (I159) NOS ULTIMOS 2A
QU2A_N083	QUANTAS VEZES FEZ O PROCEDIMENTO (N083) NOS ULTIMOS 2A
QU2A_N088	QUANTAS VEZES FEZ O PROCEDIMENTO (N088) NOS ULTIMOS 2A
QU2A_N133	QUANTAS VEZES FEZ O PROCEDIMENTO (N133) NOS ULTIMOS 2A
QU2A_N138	QUANTAS VEZES FEZ O PROCEDIMENTO (N138) NOS ULTIMOS 2A
QU2A_N139	QUANTAS VEZES FEZ O PROCEDIMENTO (N139) NOS ULTIMOS 2A
QU2A_N160	QUANTAS VEZES FEZ O PROCEDIMENTO (N160) NOS ULTIMOS 2A
QU2A_N202	QUANTAS VEZES FEZ O PROCEDIMENTO (N202) NOS ULTIMOS 2A
QU2A_N209	QUANTAS VEZES FEZ O PROCEDIMENTO (N209) NOS ULTIMOS 2A
QU2A_N210	QUANTAS VEZES FEZ O PROCEDIMENTO (N210) NOS ULTIMOS 2A
QU2A_N211	QUANTAS VEZES FEZ O PROCEDIMENTO (N211) NOS ULTIMOS 2A
QU2A_N218	QUANTAS VEZES FEZ O PROCEDIMENTO (N218) NOS ULTIMOS 2A
QU2A_N219	QUANTAS VEZES FEZ O PROCEDIMENTO (N219) NOS ULTIMOS 2A
QU2A_N433	QUANTAS VEZES FEZ O PROCEDIMENTO (N433) NOS ULTIMOS 2A
QTC_E112	A QUANTO TEMPO TEVE O CID E112
QTC_F063	A QUANTO TEMPO TEVE O CID F063

CÓDIGO	DESCRIÇÃO
QTC_F069	A QUANTO TEMPO TEVE O CID F069
QTC_F078	A QUANTO TEMPO TEVE O CID F078
QTC_F09	A QUANTO TEMPO TEVE O CID F09
QTC_F102	A QUANTO TEMPO TEVE O CID F102
QTC_F142	A QUANTO TEMPO TEVE O CID F142
QTC_F192	A QUANTO TEMPO TEVE O CID F192
QTC_F200	A QUANTO TEMPO TEVE O CID F200
QTC_F201	A QUANTO TEMPO TEVE O CID F201
QTC_F205	A QUANTO TEMPO TEVE O CID F205
QTC_F21	A QUANTO TEMPO TEVE O CID F21
QTC_F250	A QUANTO TEMPO TEVE O CID F250
QTC_F29	A QUANTO TEMPO TEVE O CID F29
QTC_F310	A QUANTO TEMPO TEVE O CID F310
QTC_F316	A QUANTO TEMPO TEVE O CID F316
QTC_F317	A QUANTO TEMPO TEVE O CID F317
QTC_F319	A QUANTO TEMPO TEVE O CID F319
QTC_F320	A QUANTO TEMPO TEVE O CID F320
QTC_F321	A QUANTO TEMPO TEVE O CID F321
QTC_F322	A QUANTO TEMPO TEVE O CID F322
QTC_F323	A QUANTO TEMPO TEVE O CID F323
QTC_F330	A QUANTO TEMPO TEVE O CID F330
QTC_F331	A QUANTO TEMPO TEVE O CID F331
QTC_F332	A QUANTO TEMPO TEVE O CID F332
QTC_F412	A QUANTO TEMPO TEVE O CID F412
QTC_F432	A QUANTO TEMPO TEVE O CID F432
QTC_F600	A QUANTO TEMPO TEVE O CID F600
QTC_H250	A QUANTO TEMPO TEVE O CID H250
QTC_H251	A QUANTO TEMPO TEVE O CID H251
QTC_H258	A QUANTO TEMPO TEVE O CID H258
QTC_H259	A QUANTO TEMPO TEVE O CID H259
QTC_H260	A QUANTO TEMPO TEVE O CID H260
QTC_H264	A QUANTO TEMPO TEVE O CID H264
QTC_H268	A QUANTO TEMPO TEVE O CID H268
QTC_H269	A QUANTO TEMPO TEVE O CID H269
QTC_H330	A QUANTO TEMPO TEVE O CID H330
QTC_H340	A QUANTO TEMPO TEVE O CID H340
QTC_H360	A QUANTO TEMPO TEVE O CID H360
QTC_H368	A QUANTO TEMPO TEVE O CID H368
QTC_H402	A QUANTO TEMPO TEVE O CID H402
QTC_H900	A QUANTO TEMPO TEVE O CID H900
QTC_H903	A QUANTO TEMPO TEVE O CID H903
QTC_H904	A QUANTO TEMPO TEVE O CID H904
QTC_H905	A QUANTO TEMPO TEVE O CID H905
QTC_H906	A QUANTO TEMPO TEVE O CID H906
QTC_H908	A QUANTO TEMPO TEVE O CID H908
QTC_H911	A QUANTO TEMPO TEVE O CID H911
QTC_H918	A QUANTO TEMPO TEVE O CID H918
QTC_H919	A QUANTO TEMPO TEVE O CID H919
QTC_H932	A QUANTO TEMPO TEVE O CID H932
QTC_I050	A QUANTO TEMPO TEVE O CID I050
QTC_I120	A QUANTO TEMPO TEVE O CID I120

CÓDIGO	DESCRIÇÃO
QTC_I159	A QUANTO TEMPO TEVE O CID I159
QTC_I200	A QUANTO TEMPO TEVE O CID I200
QTC_I201	A QUANTO TEMPO TEVE O CID I201
QTC_I208	A QUANTO TEMPO TEVE O CID I208
QTC_I209	A QUANTO TEMPO TEVE O CID I209
QTC_I210	A QUANTO TEMPO TEVE O CID I210
QTC_I219	A QUANTO TEMPO TEVE O CID I219
QTC_I220	A QUANTO TEMPO TEVE O CID I220
QTC_I248	A QUANTO TEMPO TEVE O CID I248
QTC_I249	A QUANTO TEMPO TEVE O CID I249
QTC_I250	A QUANTO TEMPO TEVE O CID I250
QTC_I251	A QUANTO TEMPO TEVE O CID I251
QTC_I255	A QUANTO TEMPO TEVE O CID I255
QTC_I258	A QUANTO TEMPO TEVE O CID I258
QTC_I259	A QUANTO TEMPO TEVE O CID I259
QTC_I442	A QUANTO TEMPO TEVE O CID I442
QTC_I443	A QUANTO TEMPO TEVE O CID I443
QTC_I498	A QUANTO TEMPO TEVE O CID I498
QTC_I499	A QUANTO TEMPO TEVE O CID I499
QTC_N083	A QUANTO TEMPO TEVE O CID N083
QTC_N088	A QUANTO TEMPO TEVE O CID N088
QTC_N133	A QUANTO TEMPO TEVE O CID N133
QTC_N138	A QUANTO TEMPO TEVE O CID N138
QTC_N139	A QUANTO TEMPO TEVE O CID N139
QTC_N160	A QUANTO TEMPO TEVE O CID N160
QTC_N200	A QUANTO TEMPO TEVE O CID N200
QTC_N201	A QUANTO TEMPO TEVE O CID N201
QTC_N202	A QUANTO TEMPO TEVE O CID N202
QTC_N209	A QUANTO TEMPO TEVE O CID N209
QTC_N210	A QUANTO TEMPO TEVE O CID N210
QTC_N211	A QUANTO TEMPO TEVE O CID N211
QTC_N218	A QUANTO TEMPO TEVE O CID N218
QTC_N219	A QUANTO TEMPO TEVE O CID N219
QTC_N433	A QUANTO TEMPO TEVE O CID N433
QTC_N47	A QUANTO TEMPO TEVE O CID N47
QTC_Z048	A QUANTO TEMPO TEVE O CID Z048
QTC_Z524	A QUANTO TEMPO TEVE O CID Z524
QTC_Z525	A QUANTO TEMPO TEVE O CID Z525
QTC_Z940	A QUANTO TEMPO TEVE O CID Z940
QTC_Z941	A QUANTO TEMPO TEVE O CID Z941
QTC_Z942	A QUANTO TEMPO TEVE O CID Z942
QTC_Z944	A QUANTO TEMPO TEVE O CID Z944
QTC_Z947	A QUANTO TEMPO TEVE O CID Z947

APÊNDICE B

Resultado das predições

Árvore de Classificação				Logístico			
Ponto de corte = 0,1		Real		Ponto de corte = 0,1		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	209.182	7.758	Predição	Saudável	322.311	9.306
	Adoeceu	196.285	41.929		Adoeceu	83.156	40.381
Ponto de corte = 0,2		Real		Ponto de corte = 0,2		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	304632	23903	Predição	Saudável	378799	18769
	Adoeceu	100835	25784		Adoeceu	26668	30918
Ponto de corte = 0,3		Real		Ponto de corte = 0,3		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	334296	33317	Predição	Saudável	392797	26288
	Adoeceu	71171	16370		Adoeceu	12670	23399
Ponto de corte = 0,4		Real		Ponto de corte = 0,4		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	337185	34713	Predição	Saudável	395641	28904
	Adoeceu	68282	14974		Adoeceu	9826	20783
Ponto de corte = 0,5		Real		Ponto de corte = 0,5		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	337405	34846	Predição	Saudável	397664	31790
	Adoeceu	68062	14841		Adoeceu	7803	17897
Ponto de corte = 0,6		Real		Ponto de corte = 0,6		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	337433	34846	Predição	Saudável	399672	35142
	Adoeceu	68034	14841		Adoeceu	5795	14545
Ponto de corte = 0,7		Real		Ponto de corte = 0,7		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	337516	34930	Predição	Saudável	401390	38477
	Adoeceu	67951	14757		Adoeceu	4077	11210
Ponto de corte = 0,8		Real		Ponto de corte = 0,8		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	337542	34930	Predição	Saudável	403301	43250
	Adoeceu	67925	14757		Adoeceu	2166	6437
Ponto de corte = 0,9		Real		Ponto de corte = 0,9		Real	
		Saudável	Adoeceu			Saudável	Adoeceu
Predição	Saudável	337575	34971	Predição	Saudável	404810	46935
	Adoeceu	67892	14716		Adoeceu	657	2752

APÊNDICE C

Modelo Logístico Ajustado

Call:

```
glm(formula = indicador ~ ., family = binomial(link = "logit"),  
     data = train_cred)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-5.9087	-0.4118	-0.2850	-0.0597	4.2227

Coefficients: (2 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept)	-3.138e+00	2.059e-02	-152.441	< 2e-16	***
uf	2.325e-02	4.262e-04	54.541	< 2e-16	***
sexoM	2.022e-01	6.744e-03	29.988	< 2e-16	***
idade	5.077e-03	2.177e-04	23.323	< 2e-16	***
qu2a	2.609e+00	3.046e-02	85.637	< 2e-16	***
qu2a_201	-2.355e+01	8.531e+02	-0.028	0.977975	
qu2a_211	-3.740e+00	4.998e-02	-74.838	< 2e-16	***
qu2a_301	-3.144e+00	3.076e-02	-102.223	< 2e-16	***
qu2a_307	-1.978e+01	2.018e+02	-0.098	0.921936	
qu2a_309	-3.758e+00	1.000e-01	-37.574	< 2e-16	***
qu2a_405	-2.959e+00	6.259e-02	-47.277	< 2e-16	***
qu2a_409	-3.410e+00	6.861e-02	-49.707	< 2e-16	***
qu2a_414	-2.219e+01	1.167e+03	-0.019	0.984831	
qu2a_503	-1.974e+01	2.916e+02	-0.068	0.946037	
qu2a_504	-1.794e+01	5.029e+02	-0.036	0.971538	
qu2a_505	-2.316e+01	1.454e+03	-0.016	0.987289	
qu2a_506	-3.720e+00	3.284e-01	-11.330	< 2e-16	***
qu2a_701	-3.261e+00	5.501e-02	-59.277	< 2e-16	***
qu2a_211020010	1.354e+00	5.320e-02	25.444	< 2e-16	***
qu2a_211070319	6.608e-01	2.793e-02	23.656	< 2e-16	***
qu2a_301070032	-6.342e-01	4.312e-02	-14.709	< 2e-16	***
qu2a_301080062	-2.075e-01	1.553e-02	-13.354	< 2e-16	***
qu2a_301080100	5.220e-02	5.900e-03	8.847	< 2e-16	***
qu2a_301080127	3.453e-02	5.981e-02	0.577	0.563725	
qu2a_301130019	1.645e-02	2.865e-02	0.574	0.565930	
qu2a_309030129	1.171e+00	1.171e-01	10.005	< 2e-16	***
qu2a_405030045	1.142e+00	1.187e-01	9.625	< 2e-16	***
qu2a_405030193	2.103e+00	1.341e-01	15.682	< 2e-16	***
qu2a_405050020	-2.771e-01	5.212e-02	-5.316	1.06e-07	***
qu2a_405050097	-4.936e-01	5.287e-02	-9.335	< 2e-16	***
qu2a_405050119	-2.681e-01	5.108e-02	-5.249	1.53e-07	***
qu2a_418020027	1.387e+00	3.288e-01	4.218	2.47e-05	***
qu2a_506010023	2.234e+00	3.759e-01	5.943	2.81e-09	***
qu2a_506010040	-5.377e-02	3.723e-02	-1.444	0.148676	
qu2a_701030127	3.784e-02	3.800e-02	0.996	0.319346	
qu2a_701030135	-1.578e+01	1.349e+03	-0.012	0.990667	

qu2a_f063	-1.577e+01	8.522e+02	-0.019	0.985232	
qu2a_f069	8.216e-01	1.455e-02	56.473	< 2e-16	***
qu2a_f078	5.167e-01	4.025e-02	12.839	< 2e-16	***
qu2a_f09	2.673e-01	2.556e-02	10.458	< 2e-16	***
qu2a_f102	1.725e-01	5.346e-02	3.226	0.001257	**
qu2a_f142	4.628e-01	2.838e-02	16.306	< 2e-16	***
qu2a_f192	4.172e-01	2.093e-02	19.934	< 2e-16	***
qu2a_f200	-1.451e+01	3.275e+02	-0.044	0.964654	
qu2a_f201	7.226e-01	3.074e-02	23.508	< 2e-16	***
qu2a_f205	-1.565e+01	5.499e+02	-0.028	0.977294	
qu2a_f21	-1.541e+01	5.552e+02	-0.028	0.977850	
qu2a_f250	-5.108e-01	6.988e-02	-7.309	2.68e-13	***
qu2a_f29	6.744e-01	2.985e-02	22.592	< 2e-16	***
qu2a_f310	-1.607e+01	7.831e+02	-0.021	0.983627	
qu2a_f316	-1.556e+01	5.295e+02	-0.029	0.976561	
qu2a_f317	-2.891e-01	6.179e-02	-4.679	2.88e-06	***
qu2a_f319	8.473e-01	2.909e-02	29.122	< 2e-16	***
qu2a_f320	5.263e-02	6.089e-02	0.864	0.387473	
qu2a_f321	2.129e-01	4.279e-02	4.975	6.53e-07	***
qu2a_f322	2.040e+00	1.090e-01	18.714	< 2e-16	***
qu2a_f323	1.707e+01	2.164e+01	0.789	0.430390	
qu2a_f330	-1.576e+01	3.240e+02	-0.049	0.961196	
qu2a_f331	-1.574e+01	4.694e+02	-0.034	0.973250	
qu2a_f332	-1.540e+01	2.260e+02	-0.068	0.945698	
qu2a_f412	-1.656e+01	7.026e+02	-0.024	0.981198	
qu2a_f432	7.199e-01	6.574e-02	10.952	< 2e-16	***
qu2a_f600	-5.088e-02	6.007e-02	-0.847	0.396998	
qu2a_h250	3.790e-01	5.819e-02	6.513	7.34e-11	***
qu2a_h251	-1.063e-01	7.154e-02	-1.486	0.137226	
qu2a_h258	4.964e-01	6.065e-02	8.185	2.72e-16	***
qu2a_h259	1.405e+00	1.188e-01	11.826	< 2e-16	***
qu2a_h260	-2.202e+00	1.493e-01	-14.753	< 2e-16	***
qu2a_h264	1.155e+00	7.912e-02	14.597	< 2e-16	***
qu2a_h268	8.242e-01	6.775e-02	12.166	< 2e-16	***
qu2a_h330	1.820e+00	1.881e-01	9.675	< 2e-16	***
qu2a_h340	1.389e-01	1.072e-01	1.295	0.195297	
qu2a_h360	6.148e-02	1.332e-01	0.462	0.644314	
qu2a_h368	6.123e-01	2.181e-01	2.808	0.004985	**
qu2a_h402	6.213e-01	5.757e-02	10.793	< 2e-16	***
qu2a_h900	2.399e-01	4.529e-02	5.297	1.18e-07	***
qu2a_h903	-1.306e+00	1.872e-01	-6.975	3.06e-12	***
qu2a_h904	3.457e-01	4.566e-02	7.572	3.67e-14	***
qu2a_h905	6.307e-01	4.992e-02	12.636	< 2e-16	***
qu2a_h906	7.169e-01	1.023e-01	7.009	2.41e-12	***
qu2a_h908	-1.014e+01	1.281e+00	-7.921	2.36e-15	***
qu2a_h911	-3.142e+02	1.321e+03	-0.238	0.812024	
qu2a_h918	3.681e-01	5.629e-02	6.539	6.19e-11	***
qu2a_h919	-8.089e-01	1.649e-01	-4.906	9.32e-07	***
qu2a_h932	2.824e+00	1.761e-01	16.038	< 2e-16	***
qu2a_i120	-8.442e-02	4.611e-02	-1.831	0.067152	.

qu2a_i200	-2.189e+01	1.265e+04	-0.002	0.998619
qu2a_i201	-4.291e-01	9.972e-02	-4.303	1.68e-05 ***
qu2a_i208	-1.086e-01	4.447e-02	-2.441	0.014631 *
qu2a_i209	-2.129e+01	4.446e+03	-0.005	0.996179
qu2a_i210	-5.372e-01	7.179e-02	-7.483	7.28e-14 ***
qu2a_i219	-2.219e+01	7.567e+03	-0.003	0.997660
qu2a_i220	-6.148e-01	1.050e-01	-5.853	4.82e-09 ***
qu2a_i248	-1.472e+00	1.284e-01	-11.466	< 2e-16 ***
qu2a_i249	-2.229e+00	1.286e-01	-17.333	< 2e-16 ***
qu2a_i250	4.281e-02	1.264e-01	0.339	0.734921
qu2a_i251	-2.438e+01	7.716e+03	-0.003	0.997479
qu2a_i255	-2.192e+01	9.507e+03	-0.002	0.998160
qu2a_i258	4.281e-02	5.630e-02	0.760	0.447043
qu2a_i259	-8.863e-02	8.605e-02	-1.030	0.303016
qu2a_i442	5.400e+00	4.121e-01	13.105	< 2e-16 ***
qu2a_i443	-2.159e+00	4.562e-01	-4.733	2.21e-06 ***
qu2a_i498	4.899e-01	1.047e-01	4.681	2.86e-06 ***
qu2a_i499	5.996e-01	9.908e-02	6.052	1.43e-09 ***
qu2a_n200	1.143e+00	1.334e-01	8.570	< 2e-16 ***
qu2a_n201	-1.998e+01	4.231e+03	-0.005	0.996233
qu2a_z048	2.502e+00	4.093e-01	6.113	9.80e-10 ***
qu2a_z524	-1.889e+01	3.748e+02	-0.050	0.959811
qu2a_z525	-2.422e-01	4.080e-02	-5.936	2.92e-09 ***
qu2a_z940	-1.944e+01	1.973e+03	-0.010	0.992137
qu2a_z941	-1.741e+01	4.652e+03	-0.004	0.997013
qu2a_z942	-4.790e-01	4.805e-02	-9.969	< 2e-16 ***
qu2a_z944	-1.718e-01	3.358e-01	-0.512	0.608836
qu2a_z947	-5.191e-01	4.656e-02	-11.151	< 2e-16 ***
qu2a_n088	-1.604e+01	1.989e+05	0.000	0.999936
qu2a_n160	-2.041e+01	8.176e+03	-0.002	0.998008
qu2a_n202	-1.727e+01	2.607e+00	-6.627	3.42e-11 ***
qu2a_n209	-1.923e+01	6.706e+03	-0.003	0.997713
qu2a_n210	-1.949e+01	8.989e+03	-0.002	0.998270
qu2a_n211	-2.040e+01	9.367e+03	-0.002	0.998262
qu2a_n218	-1.911e+01	2.651e+04	-0.001	0.999425
qu2a_n219	-2.041e+01	2.799e+04	-0.001	0.999418
qtc_e112	7.246e-03	8.198e+01	0.000	0.999929
qtc_f063	1.101e-02	5.121e+01	0.000	0.999828
qtc_f069	-9.178e-03	9.741e-04	-9.422	< 2e-16 ***
qtc_f078	1.535e-02	3.095e-03	4.959	7.09e-07 ***
qtc_f09	4.610e-03	1.802e-03	2.559	0.010505 *
qtc_f102	2.305e-02	3.248e-03	7.096	1.29e-12 ***
qtc_f142	-1.316e-02	2.225e-03	-5.914	3.33e-09 ***
qtc_f192	1.696e-03	1.525e-03	1.112	0.266151
qtc_f200	2.521e-03	2.070e+01	0.000	0.999903
qtc_f201	-1.839e-02	2.657e-03	-6.922	4.45e-12 ***
qtc_f205	6.295e-02	3.481e+01	0.002	0.998557
qtc_f21	2.022e-02	3.512e+01	0.001	0.999540
qtc_f250	5.569e-02	3.745e-03	14.871	< 2e-16 ***
qtc_f29	-9.567e-03	2.322e-03	-4.121	3.78e-05 ***

qtc_f310	-5.387e-03	4.864e+01	0.000	0.999912
qtc_f316	1.427e-02	3.331e+01	0.000	0.999658
qtc_f317	6.310e-02	3.615e-03	17.459	< 2e-16 ***
qtc_f319	-5.786e-02	3.604e-03	-16.053	< 2e-16 ***
qtc_f320	-1.603e-02	5.143e-03	-3.116	0.001833 **
qtc_f321	1.413e-02	2.870e-03	4.925	8.45e-07 ***
qtc_f322	-4.091e-01	3.855e-02	-10.612	< 2e-16 ***
qtc_f323	-1.418e+01	2.164e+01	-0.655	0.512221
qtc_f330	2.014e-02	2.072e+01	0.001	0.999224
qtc_f331	1.018e-02	2.977e+01	0.000	0.999727
qtc_f332	2.978e-03	1.434e+01	0.000	0.999834
qtc_f412	2.562e-02	4.393e+01	0.001	0.999535
qtc_f432	3.602e-03	5.071e-03	0.710	0.477519
qtc_f600	2.022e-02	2.109e-03	9.587	< 2e-16 ***
qtc_h250	7.333e-04	1.938e-03	0.378	0.705079
qtc_h251	-5.293e-03	3.675e-03	-1.440	0.149870
qtc_h258	-1.362e-02	2.272e-03	-5.995	2.03e-09 ***
qtc_h259	-8.757e-02	6.161e-03	-14.214	< 2e-16 ***
qtc_h260	-1.241e-02	4.467e-03	-2.779	0.005451 **
qtc_h264	-1.372e-02	4.392e-03	-3.123	0.001791 **
qtc_h268	-1.069e-01	4.528e-03	-23.601	< 2e-16 ***
qtc_h330	-1.802e-01	1.459e-02	-12.350	< 2e-16 ***
qtc_h340	9.167e-03	1.323e-03	6.927	4.29e-12 ***
qtc_h360	3.611e-03	4.666e-03	0.774	0.438963
qtc_h368	-2.986e-02	1.438e-02	-2.076	0.037883 *
qtc_h402	-1.488e-02	3.603e-03	-4.131	3.61e-05 ***
qtc_h900	-7.010e-03	1.892e-03	-3.706	0.000211 ***
qtc_h903	1.262e-01	9.677e-03	13.044	< 2e-16 ***
qtc_h904	1.310e-02	1.791e-03	7.313	2.60e-13 ***
qtc_h905	3.689e-03	2.312e-03	1.596	0.110464
qtc_h906	-2.433e-02	7.083e-03	-3.436	0.000591 ***
qtc_h908	4.710e-01	5.538e-02	8.505	< 2e-16 ***
qtc_h911	1.318e+01	5.505e+01	0.239	0.810765
qtc_h918	-9.754e-03	3.198e-03	-3.050	0.002286 **
qtc_h919	-1.016e-02	1.230e-02	-0.826	0.408968
qtc_h932	-1.464e-01	1.361e-02	-10.755	< 2e-16 ***
qtc_i159	3.819e-02	2.003e-03	19.064	< 2e-16 ***
qtc_i200	-3.632e-02	9.404e+02	0.000	0.999969
qtc_i201	5.015e-02	5.990e-03	8.373	< 2e-16 ***
qtc_i208	1.546e-02	2.088e-03	7.403	1.34e-13 ***
qtc_i209	-6.770e-02	3.719e+02	0.000	0.999855
qtc_i210	6.096e-02	4.364e-03	13.968	< 2e-16 ***
qtc_i220	-1.830e-02	8.053e-03	-2.273	0.023040 *
qtc_i248	7.994e-02	7.987e-03	10.008	< 2e-16 ***
qtc_i249	9.743e-02	7.111e-03	13.701	< 2e-16 ***
qtc_i250	1.188e-02	7.793e-03	1.524	0.127540
qtc_i251	9.417e-02	4.824e+02	0.000	0.999844
qtc_i255	-4.119e-03	7.237e+02	0.000	0.999995
qtc_i258	-1.933e-02	3.616e-03	-5.347	8.96e-08 ***
qtc_i259	4.220e-02	3.963e-03	10.647	< 2e-16 ***

```

qtc_i442    -2.685e-01  4.053e-02  -6.624 3.49e-11 ***
qtc_i443    2.227e-01  2.292e-02   9.716 < 2e-16 ***
qtc_i498    -2.385e-02  7.966e-03  -2.994 0.002750 **
qtc_n088    -2.367e-01  1.162e+04   0.000 0.999984
qtc_n160    3.008e-02  2.265e-03  13.278 < 2e-16 ***
qtc_n200    2.346e-02  6.365e-03   3.686 0.000228 ***
qtc_n201    2.470e-02  5.416e+02   0.000 0.999964
qtc_n202    8.735e-01  1.145e-01   7.628 2.38e-14 ***
qtc_n209    1.129e-01  3.985e+02   0.000 0.999774
qtc_n210    2.087e-02  6.719e+02   0.000 0.999975
qtc_n211    3.542e-02  6.296e+02   0.000 0.999955
qtc_n218    2.298e-02  1.603e+03   0.000 0.999989
[ reached getOption("max.print") -- omitted 11 rows ]

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 941518 on 1365458 degrees of freedom
Residual deviance: 637245 on 1365250 degrees of freedom
AIC: 637663

Number of Fisher Scoring iterations: 23