UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

JANAÍNA APARECIDA CEZÁRIO

UTILIZAÇÃO DO ALGORITMO APRIORI E REGRESSÃO LOGÍSTICA EM UM ESTUDO SOBRE A CESSAÇÃO DO TABAGISMO

CAMPINA GRANDE - PB

FEVEREIRO 2020

JANAÍNA APARECIDA CEZÁRIO

UTILIZAÇÃO DO ALGORITMO APRIORI E REGRESSÃO LOGÍSTICA EM UM ESTUDO SOBRE A CESSAÇÃO DO TABAGISMO

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: KLEBER NAPOLEÃO NUNES DE OLIVEIRA BARROS

CAMPINA GRANDE - PB

FEVEREIRO 2020

C425u Cezário, Janaína Aparecida.

Utilização do Algoritmo Apriori e Regressão logística em um estudo sobre a cessação do tabagismo [manuscrito] / Janaina Aparecida Cezario. - 2020.

45 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2020.

"Orientação : Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros , Departamento de Estatística - CCT."

1. Algoritmo Apriori. 2. Regressão logística. 3. Tabagismo. I. Título

21. ed. CDD 519.5

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

JANAÍNA APARECIDA CEZÁRIO

UTILIZAÇÃO DO ALGORITMO APRIORI E REGRESSÃO LOGÍSTICA EM UM ESTUDO SOBRE A CESSAÇÃO DO TABAGISMO

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 28 de Fevereiro de 2020.

BANCA EXAMINADORA

Prof. Dr Kleber Napoleão N. de O. Barros
Universidade Estadual da Paraíba
(Orientador)

Prof. Dr. Mácio Augusto de Albuquerque
Universidade Estadual da Paraíba

Prof. Dr. Gilberto da Silva Matos
Universidade Federal de Campina Grande

Dedico este trabalho aos meus pais, Rosângela e João, minha fortaleza e incentivo de caminhar um pouco mais. Obrigada por tudo, vocês sempre vão ser minha fonte de energia necessária para prosseguir a passos largos nesse caminho chamado vida.

Agradecimentos

Ao Grande e Todo poderoso **DEUS** seja toda Honra e toda Glória, para Ele a minha gratidão eterna por tantos momentos que somente Ele estava comigo e isso eu não terei como pagar tudo quanto foi feito por mim.

Aos meus familiares e amigos, a gratidão por tantos momentos de compreensão e de incentivo onde pude compreender realmente o que é a palavra amor.

Ao meu orientador Profº Dr. Kleber Napoleão meu professor por quase o curso inteiro, a minha gratidão por tudo quanto foi feito por mim, toda ajuda, todo ensinamento, até nos momentos de raiva eu pude aprender e crescer profissionalmente. Ganhei não apenas um professor, mais sim um amigo e colega de profissão que levarei por toda minha vida.

Aos meus professores de curso, minha motivação sempre vão ser vocês pois tudo se torna simples com as pessoas certas nos lugares certos, e assim eram vocês para mim a cada semestre cursado. Obrigada por tanto conhecimento, por tantas portas abertas e por tanto amor por essa ciência incrível.

Por fim, a todos que me ajudaram de maneira direta ou indiretamente para que chegasse até aqui minha eterna gratidão.

"Nada é tão doce quanto descansar nas mãos de Deus, e saber apenas a sua vontade."
(Charles H. Spurgeon)

Resumo

Em nossos dias atuais descobrir padrões, facilitar a vida das pessoas é sempre um desafio para quem se trabalha com banco de dados além de extrair informações importantes e cruciais que venham a modificar a vida das pessoas que necessitam de tais informações.A compra de um certo produto que tem relação com a compra de outro produto representa uma regra de associação, regras essas que são utilizadas em diversas áreas. Para descobrir tais regras, em dados categóricos existe a utilização do algoritmo Apriori que com essa associação tem como objetivo encontrar elementos que implicam na presença de outros elementos em uma mesma transação de um banco de dados. A importância de se trabalhar com modelos lineares generalizados é de que os dados não podem ser ajustados com apenas uma regressão linear simples. Neste trabalho apresentamos um ajuste de modelo de regressão logística com o auxilio do algoritmo Apriori que foram aplicados a dados obtidos do Hospital Universitário Alcides Carneiro localizado na cidade de Campina Grande-PB. As informações se referem a pacientes atendidos pelo programa Multidisciplinar de tratamento do tabagismo vinculado ao curso de Medicina da Universidade Federal de Campina Grande. Como objetivo principal temos a identificação via modelo de regressão logística de padrões de usuários com a maior probabilidade de cessação do tabagismo a partir de variáveis filtradas pelo algoritmo Apriori. Para a limpeza dos dados foi utilizado o programa `Open Refine` e o ajuste diagnóstico do modelo foi realizado com auxilio do programa `R`. Constatou-se que as variáveis que compõe o modelo final são o Estado Civil e Religião.

Palavras-chaves: Algoritmo Apriori. Regressão Logística. Cessação do tabagismo.

Abstract

Nowadays discover patterns that can make people's lives easier is always a challenge for those who work with databases, in addition to extracting importante and crucial information that will change the lives of people who need this information. The purchase of a certain product that is related to the purchase of another product representes a rule of association, rules that are used in several areas. To find these rules, in categorical data there is the use of the Apriori algorithm which with this association has like objective to find elements that imply the presence of other elements in the same database transaction. The importance of working with generalized linear models is that the data cannot be adjusted with just a simple linear regression. This work presents an adjustment of the logistic regression model with the aid of the Apriori algorithm that were applied to data obtained at the university hospital Alcides Carneiro located in Campina Grande city. The information refers to patients treated by the multidisciplinar smoking treatment program linked to the medical course at the Federal University of Campina Grande. The main objective was to identify, through the logistic regression model, user patterns with the highest probability of smoking cessation from variables filtered by the Apriori algorithm. For data cleaning, the Open Refine program was used and the model's diagnostic adjustment was carried out with the aid of the R. program. It was found that the variable that make up the final model are marital status and religion.

Key-words: Apriori algorithm, logistic regression and smoking cessation.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A Organização Mundial da Saúde (OMS) estima que mais de cinco milhões de mortes ao ano no mundo são decorrentes do tabagismo e espera-se que esse número seja de aproximadamente oito milhões no ano 2030, tornando o tabagismo a principal causa de morte prematura, sendo que 80% delas ocorrerão em países em desenvolvimento, um número alto e alarmante que só tende a crescer segundo diversas pesquisas realizadas no Brasil e no mundo (ORGANIZATION; CONTROL, 2008).

No Brasil, o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) é o órgão do Ministério da Saúde responsável pelo Programa Nacional de Controle do Tabagismo (PNCT) e pela articulação da rede de tratamento do tabagismo no SUS (Sistema Único de Saúde), em parceria com estados e municípios e Distrito Federal. Assim o tratamento de tabagismo é desenvolvido segundo a base nas diretrizes do PNCT, que está sob a coordenação e gerenciamento da Divisão de Controle do Tabagismo e Outros Fatores de Risco (DITAB), do INCA. As ações educativas, legislativas e econômicas desenvolvidas vêm gerando uma diminuição da aceitação social do tabagismo, fazendo com que um número cada vez maior de pessoas queiram parar de fumar, evidenciando assim a importância de se priorizar o tratamento do fumante como uma estratégia fundamental no controle do tabagismo (INCA, 2017).

Mesmo com a segunda maior produção e sendo o maior exportador de tabaco, o Brasil tem conseguido desenvolver ações para controlar o tabagismo, ações fortes e abrangentes, o que tem lhe conferido o reconhecimento de liderança internacional nessa área (CAVALCANTE, 2005).

A partir do ano de 2003 o Ministério da Saúde através de sua Secretaria de Vigilância em Saúde (SVS) passou a estruturar um Sistema Nacional de Vigilância específico para as doenças não transmissíveis e seus fatores de risco, dentre eles temos o tabagismo. Nesse mesmo ano, o INCA em parceria com a SVS desenvolveu o Inquérito Domiciliar Sobre Comportamentos de Risco e Morbidade Referida de Doenças e Agravos Não Transmissíveis em 15 capitais brasileiras e Distrito Federal, e em 2008 participou ativamente da Pesquisa Especial sobre Tabagismo (PETAB), coordenada pelo Ministério da Saúde e IBGE – quando o Brasil aderiu ao Global Adult Tobacco Survey (GATS) proposto pela OMS e Centers for Disease Control and Prevention (CDC). Anualmente, R$ 56,9 bilhões são gastos pelo Brasil com despesas médicas e em perda de produtividade provocadas pelo tabagismo. Em compensação, o país arrecada por ano apenas R$ 13 bilhões em impostos sobre a venda de cigarros, ou seja, esse valor cobre apenas 23% dos gastos com os males causados pela epidemia do tabaco (INCA, 2017).

Atualmente, a sociedade tem atribuído grande importância a diversas informações transmitidas pelos meios de comunicação, as quais, na maioria das vezes, vêm expressas por listas, tabelas e gráficos de vários tipos. Nesse sentido, é importante que tenhamos os conhecimentos e ferramentas necessárias para entendermos o significado desses dados e, ao mesmo tempo, que saibamos interpretar os diferentes instrumentos que são utilizados para representá-los. Por outro lado, para desenvolver essa capacidade não só de entender o argumento apresentado mas, também de criticá-lo, é importante saber selecionar, organizar e entender as informações, que nos são mostradas a todo o momento, nisso compreendemos a importância de se utilizar da Estatística descritiva. A Estatística descritiva visa a sumarizar e descrever qualquer conjunto de dados. Em outras palavras, é aquela estatística que está preocupada em sintetizar os dados de uma maneira direta, preocupando-se menos com variações e intervalos de confiança dos dados.

Avaliar um banco de dados nem sempre é uma tarefa fácil, temos de observar desde o tamanho da amostra, o número de variáveis, as variáveis que serão necessárias para ser realizado o estudo no que está sendo desenvolvido e entre outros. Sabendo que não basta apenas obter os dados mas se faz necessário também fazer a limpeza desses dados. Para isso, utilizaremos o `Open Refine` que é um programa que possibilita o tratamento e a manipulação de dados, especialmente quando estes estão desorganizados ou apresentam inconsistências. Suas funções vão desde limpar, corrigir, clusterizar e filtrar os dados, até transformá-los de um formato para outro. Para utilizá-lo, é necessário fazer o seu download gratuitamente no site do software (https://openrefine.org/download.html). Nesse mesmo site, também é possível encontrar tutoriais que explicam melhor como o programa funciona. O surgimento do `Open Refine` se deu com a criação do Freebase Gridworks pela empresa Metaweb Technologies, Inc. Em sua primeira versão, ele era uma ferramenta para limpeza e upload de dados. Com a aquisição da Metaweb pelo Google, em 2010, o programa foi renomeado para Google Refine. Foram realizadas três atualizações com o apoio de engenheiros do Google, mas, desde 2012, a gigante das buscas não apoia mais o projeto – que, a partir daí, passou a ser conhecido como `Open Refine`. (SCOGNAMIGLIO, 2016).

Além da descrição e limpeza dos dados, trazemos também a técnica de associação que visa classificar um padrão de itens em uma base de dados de uma aplicação que ocorre com uma certa frequência (TARGA, 2002). Um dos algoritmos mais referenciados para realizar a tarefa de associação é o Apriori, utilizado no trabalho proposto. Ele avalia e retorna associações relevantes entre os itens, segundo critérios de suporte e confiança.

Os modelos de regressão logística são comumente utilizados em problemas de classificação, em situações que a variável dependente é de natureza dicotômica ou binária, onde as variáveis independentes podem ser categóricas ou não. Através dessa técnica, é possível também estimar a probabilidade associada à ocorrência de um determinado evento em função do conjunto de variáveis explicativas (FREITAS, 2019).

Esse trabalho tem como objetivo principal a identificação via modelo de regressão logística de padrões de usuários com maior probabilidade de cessação do tabagismo a partir das variáveis filtradas pelo algoritmo Apriori, e temos como objetivos secundários realizar a limpeza das variáveis dispostas na base de dados, descobrir regras de sequência a partir de uma base de dados real e por fim aplicar a regressão logística e selecionar o melhor modelo que se adeque aos dados analisados.

2 Fundamentação Teórica

2.1 Tabagismo no Brasil

O tabagismo é a principal causa de morte prematura no mundo. O hábito de fumar reduz a expectativa de vida média em 10 anos. Sua interrupção antes dos 40 anos reduz o risco de morte em até 90%. Dependendo da idade da cessação, a expectativa de vida pode aumentar entre quatro e dez anos. Diversas condições clínicas estão associadas ao tabagismo ou são complicadas pelo mesmo, com destaque para: doença pulmonar obstrutiva crônica, asma, doença pulmonar intersticial, doença cardiovascular, osteoporose e câncer. Considerando a magnitude do problema e sua vulnerabilidade a tecnologias leves, é indiscutível o papel da atenção primária no manejo da condição (REGULASUS, 2015).

No Brasil são plantados os tipos de fumo Virgínia (81%), Burley (17%), comum (0,8%) e outros (1,2%), onde se encontram os fumos para capa de charuto, oriental e fumo em corda. Na fabricação do cigarro são usados 40% de fumo Virgínia, 35% de fumo Burley, 15% de fumo Oriental e 10% de talo picado. A mistura destes tipos de fumo na composição do cigarro produz perfeito equilíbrio no sabor e aroma, atendendo a exigências do mercado consumidor (KIST et al., 2004).

No Brasil, observou-se uma redução de 35% na percentagem de fumadores no período de 1989 a 2003 (MONTEIRO et al., 2007), o que coincide com a adoção de campanhas educativas e leis antitabagismo mais numerosas e incisivas, incluindo a proibição da propaganda através dos meios de comunicação. No entanto, 178.000 mortes anuais em adultos com 35 ou mais anos foram atribuídas ao tabagismo em 2003 (DANTAS et al., 2017).

2.2 Hospital Alcides Carneiro

O Hospital Universitário Alcides Carneiro - HUAC é uma instituição Federal de atendimento hospitalar vinculada à Universidade Federal de Campina Grande, situada no município de Campina Grande - PB, que atende casos de alta e média complexidade, que possui 35 serviços ambulatoriais e 14 serviços de apoio, diagnóstico e tratamento. O HUAC serve de referência para mais de 180 cidades do interior paraibano, além de outros estados para pacientes com diversas condições.

2.3 Programa Multidisciplinar de Tratamento do Tabagismo

Iniciado em 2007, o Programa Multidisciplinar de Tratamento do Tabagismo (PMTT), vinculado ao Curso de Medicina do CCBS, atende por ano cerca de 400 usuários tabagistas do município de Campina Grande e circunvizinhança. A equipe é multiprofissional, multidisciplinar e interdisciplinar, composta por docentes e discentes de diversos cursos da área de saúde da UFCG e de outras universidades parceiras. Primeiramente, cada grupo de cerca de 30 novos usuários assistem a uma palestra inicial, que abrange os males associados ao tabagismo e explica todo o funcionamento do PMTT. Posteriormente, cada um deles é entrevistado e avaliado pelas equipes que integram o Programa, recebendo assistência multiprofissional. A partir de indicação médica, os usuários recebem o medicamento bupropiona, em quantidade suficiente para 15 dias, quando serão reavaliados por todas as equipes, nos retornos em grupo. O tratamento dura 12 semanas (6 encontros denominados de retornos), sendo os usuários acompanhados e incentivados a participarem de todas as atividades. Os dados da entrevista são anotados no Questionário de Avaliação Médica do Paciente. Os dados dos retornos são registrados na Ficha de Retorno.

2.4 Algoritmo Apriori

Diferentes algoritmos estatísticos foram desenvolvidos para implementar a mineração de regras de associação e o algoritmo Apriori é um desses. O algoritmo Apriori é um algoritmo que encontra regras de associação, sem se preocupar com a ordem temporal dos itens destas regras. Pra encontrar as regras, é realizado um procedimento iterativo, onde cada iteração executa duas ações: geração de candidatos possivelmente frequentes, e definição dos padrões frequentes. Ao final da execução, possuímos regras destes itens que são frequentes.

A regra de associação está na forma de A ⇒ B, onde A e B são dois conjuntos de itens separados. As três medidas amplamente usadas para selecionar regras interessantes são: suporte, confiança e lift. O *suporte* é a porcentagem de casos nos dados que contêm ambos A e B, *confiança* é a porcentagem de casos que contêm A e também B, *lift* é a razão de confiança para a porcentagem de casos que contêm B. As fórmulas para calculá-los são:

$$
\begin{aligned}
&Suporte(A \Rightarrow B) = P(A \cap B) \\
&Confiança(A \Rightarrow B) = P(A|B) \\
&= \frac{P(A\ U\ B)}{P\ (A)} \\
&Lift\ (A \Rightarrow B) = \frac{Confiança\ (A \Rightarrow B)}{P\ (B)} \\
&= \frac{P(A\ U\ B)}{P(A)P(B)}
\end{aligned}
\tag{2.1}
$$

$P(A)$ é a frequência (ou probabilidade) de casos que contêm A (ZHAO, 2012).

Para grandes conjuntos de dados, podem haver centenas de itens em centenas de milhares de linhas. O algoritmo Apriori tenta extrair regras para cada combinação possível de itens (MALIK, 2018).

Para um melhor entendimento do que descrevemos acima, segue abaixo um exemplo simples e prático de como podemos aplicar a regra de associação.

Dado a Figura 1 abaixo onde cada registro corresponde a uma transação de um cliente, com itens assumindo valores binários (sim/não), indicando se o cliente comprou ou não o respectivo item, descobrir todas as regras associativas com suporte $\geq$ 0,3 e grau de certeza (confiança) $\geq$ 0,8.

TID	leite	café	cerveja	pão	manteiga	arroz	feijão
1	não	sim	não	sim	sim	não	não
2	sim	não	sim	sim	sim	não	não
3	não	sim	não	sim	sim	não	não
4	sim	sim	não	sim	sim	não	não
5	não	não	sim	não	não	não	não
6	não	não	não	não	sim	não	não
7	não	não	não	sim	não	não	não
8	não	não	não	não	não	não	sim
9	não	não	não	não	não	sim	sim
10	não	não	não	não	não	sim	não

Figura 1 – Registro de compras de clientes.

Dada uma regra de associação “Se compra X então compra Y”, os fatores sup e conf são:

$$sup = \frac{\text{Número de registros com X e Y}}{\text{Número total de registros}} \tag{2.2}$$

$$conf = \frac{\text{Número de registros com X e Y}}{\text{Número de registros com X}} \tag{2.3}$$

Essas regras nos confirmam o que já tinham sido descritos anteriormente, analisaremos agora os dados segundo a Figura 2, seguiremos cumprindo com três passos a análise desses dados. O primeiro passo é encontrar o suporte de conjuntos com um item determinando os itens frequentes com sup $\geq$ 0,3:

Temos então nessa Figura 3 o resultado do primeiro passo, ou seja, verificamos o suporte e aqueles que seguem a regra sugerida é a que segue para o segundo passo. Passaremos ao segundo passo como nos indica a Figura 4, que é calcular o suporte de conjuntos com dois itens frequentes e com sup $\geq$ 0,3, lembrando que se um item não é frequente no primeiro passo ele pode ser ignorado aqui:

Conjunto de itens	suporte
{leite}	2
{café}	3
{cerveja}	2
{pão}	5
{manteiga}	5
{arroz}	2
{feijão}	2

Figura 2 – Suporte de conjuntos com um item

Conjunto de itens	suporte
{café}	3
{pão}	5
{manteiga}	5

Figura 3 – Conjunto resultante do primeiro passo

Conjunto de itens	suporte
{café, pão}	3
{café, manteiga}	3
{pão, manteiga}	4

Figura 4 – Conjunto resultante do segundo passo

Agora iremos ao terceiro passo, ressaltando que só é necessário considerar conjuntos de itens que são frequentes no passo anterior, com isso determinaremos conjuntos de itens frequentes com sup $\geq$ 0,3, vejamos a Figura 5:

Conjunto de itens	suporte
{café, pão, manteiga}	3

Figura 5 – Conjunto resultante do terceiro passo

Agora calcularemos regras candidatas com dois itens com o seu valor de certeza:

- Conjunto de itens:café, pão

Itens	Confiança
Se compra café então compra pão	1,0
Se compra pão então compra café	0,6

- Conjunto de itens: café, manteiga.

Itens	Confiança
Se compra café então compra manteiga	1,0
Se compra manteiga então compra café	0,6

- Conjunto de itens: pão,manteiga

Itens	Confiança
Se compra pão então compra manteiga	0,8
Se compra manteiga então compra pão	0,8

Regras candidatas com três itens com o seu valor de certeza:

- Conjunto de itens: café,manteiga,pão.

Itens	Confiança
Se compra café,manteiga então compra pão	1,0
Se compra café,pão então compra manteiga	1,0
Se compra manteiga, pão então compra café	0,75
Se compra café então compra manteiga, pão	1,0
Se compra manteiga então compra café, pão	0,6
Se compra pão então compra café, manteiga	0,6

Temos então os padrões descobertos, minsup = 0,3 e minconf = 0,8:

Itens	Suporte	Confiança
Se compra café então compra pão	0,3	1,0
Se compra café então compra manteiga	0,3	1,0
Se compra pão então compra manteiga	0,4	0,8
Se compra manteiga então compra pão	0,4	0,8
Se compra café, manteiga então compra pão	0,3	1,0
Se compra café, pão então compra manteiga	0,3	1,0
Se compra café então compra manteiga, pão	0,3	1,0

Essas são as regras encontradas que atendem ao suporte e a confiança pedidas.

2.5 Regressão Logística

Antes de iniciar a definição de Regressão Logística, devemos primeiramente comentar um pouco sobre os Modelos Lineares Generalizados (MLG). Nelder e Wedderburn apresentaram em 1972 uma proposta revolucionária, onde, unia as teorias de modelagem estatística por meio de uma classe de modelos de regressão, intitulada de modelos lineares generalizados (MLG). Nos MLG a ideia é de adequar a regressão linear a vários tipos de dados, ou seja, variar as opções para a distribuição da variável resposta, possibilitando que ela pertença à família exponencial de distribuições, assim como também flexibilizar a relação operacional entre o preditor linear e o valor esperado da variável resposta (POSSAMAI, 2018).

A estrutura geral do modelo linear normal em alguns casos se tornava inadequado por não atender as premissas de sua aplicação. Por exemplo, os erros ε não eram identicamente distribuídos como uma normal ou não eram independentes, ou ainda, ocorria à violação da suposição de homogeneidade da variância desses erros. Considerando essa situação, que Nelder e Wedderburn Nelder e Wedderburn (1972) unificaram todos os modelos apresentados anteriormente e o chamou de Modelos Lineares Generalizados MLG (DEMETRIO, 2002) .

Tal como neste estudo, quando a variável resposta Y for uma variável qualitativa e assumir dois valores, seguindo uma distribuição Bernoulli, ela faz parte de um dos casos do MLG, onde pretendemos modelar a variável resposta categórica com dois valores possíveis dado um conjunto de variáveis explicativas.

A distribuição Bernoulli se associa com a regressão logística binária devido a esse modelo de regressão buscar a caracterização apenas do "sucesso" e "fracasso" do público, podendo ser representadas pelos valores 1 e 0, respectivamente (POSSAMAI, 2018).

Podemos dizer então que a Regressão Logística é definida como uma técnica estatística que possibilita a estimação da probabilidade de ocorrência de um determinado evento em presença de um conjunto de variáveis explicativas, como também ajudar na classificação dos objetos ou casos (CORRAR et al., 2011).

O modelo logístico acabou se tornando um dos principais métodos de modelagem estatística devido à facilidade de interpretação dos parâmetros. Este fato pode ser observado através de estudos em que os pesquisadores têm dicotomizado a resposta de dados que originalmente não são binários, para modelar a probabilidade de sucesso por meio da regressão logística, como acontece em análise de sobrevivência discreta (PAULA, 2004).

Considere inicialmente o modelo logístico linear simples que nos diz (PAULA, 2004) em que $\pi(x)$, a probabilidade de sucesso dado o valor x de uma variável explicativa

qualquer é definida tal que:

$$log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \alpha + \beta x \tag{2.4}$$

em que α e β são parâmetros desconhecidos. Esse modelo poderia, por exemplo, ser aplicado para analisar a associação entre uma determinada doença e a ocorrência ou não de um fator particular. Seriam então amostrados, independentemente, n_1 indivíduos com presença do fator ($x = 1$) e n_2 indivíduos com ausência do fator ($x = 0$) e $\pi(x)$ seria a probabilidade de desenvolvimento da doença após um certo período fixo. Dessa forma, a chance de desenvolvimento da doença para um indivíduo com presença do fator fica dada por:

$$\frac{\pi(1)}{1-\pi(1)} = e^{\alpha+\beta} \tag{2.5}$$

enquanto que a chance de desenvolvimento da doença para um indivíduo com ausência do fator é simplesmente

$$\frac{\pi(0)}{1-\pi(0)} = e^{\alpha} \tag{2.6}$$

Logo, a razão de chances fica dada por:

$$\psi = \frac{\pi(1)\left\{1-\pi(0)\right\}}{\pi(0)\left\{1-\pi(1)\right\}} = e^{\beta} \tag{2.7}$$

dependendo apenas do parâmetro β. Mesmo que a amostragem seja retrospectiva, isto é, são amostrados n_1 indivíduos doentes e n_2 indivíduos não doentes, o resultado acima continua valendo. Essa é uma das grandes vantagens da regressão logística, a possibilidade de interpretação direta dos coeficientes como medidas de associação. Esse tipo de interpretação pode ser estendido para qualquer problema prático (PAULA, 2004).

2.6 Seleção de Modelos

Para a seleção de modelos existem vários procedimentos, iremos descrever alguns deles a seguir.

2.6.1 Método forward

Nesse método ajustamos um modelo somente com seu intercepto, ou seja, $\mu_{ij} = \beta_0$. Ajustamos então para cada variável explicativa o seguinte modelo:

$$\mu_{ij} = \beta_0 + \beta_j x_{ij}, \quad (j = 1, ..., p-1) \tag{2.8}$$

Testamos

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Sendo P o menor nível descritivo dentre os q testes. Se $P \leq P_E$, a variável correspondente entra no modelo. Vamos supor que X_1 tenho sido escolhida. Então, no passo seguinte ajustamos os modelos

$$\mu_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_j x_{ij}, \quad (j = 2, ..., p-1) \tag{2.9}$$

Testamos

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Seja P o menor nível descritivo dentre os *(q - 1)* testes. Se $P \leq P_E$, a variável correspondente entra no modelo. Repetimos o procedimento até que ocorra $P > P_E$.

2.6.2 Método backward

Iniciamos o procedimento pelo modelo

$$\mu_{ij} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2.10}$$

Testamos

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

para $j = 1, ..., p-1$. Seja P o maior nível descritivo dentre os q testes. Se $P > P_S$, a variável correspondente sai do modelo. Suponha que X_1 tenho saído do modelo. Então, ajustamos o modelo

$$\mu_{ij} = \beta_0 + \beta_2 x_2 + ... + \beta_p x_p \tag{2.11}$$

Testamos

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

para $j = 2, ..., p-1$. Seja P o maior nível descritivo dentre os (q - 1) testes. Se $P > PS$, então a variável correspondente sai do modelo. Repetimos o procedimento até que ocorra $P \leq P_S$.

2.6.3 Método stepwise

É uma mistura dos dois procedimentos anteriores. Iniciamos o processo com o modelo $\mu = \alpha$. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira não sai do modelo. O processo continua até que nenhuma variável seja incluída ou seja retirada do modelo. Geralmente adotamos 0, $15 \leq PE, PS \leq 0,25$. Uma sugestão seria usar $PE = PS = 0,20$. Podemos dizer então stepwise é uma ferramenta automática usada nos estágios exploratórios da construção de modelos para identificar um subconjunto útil de preditores. O processo adiciona sistematicamente a variável mais significativa ou remove a variável menos significativa durante cada etapa (PAULA, 2004).

2.6.4 Critério de Informação de Akaike

O Critério de Informação de Akaike - AIC (Akaike's Information Criterion) procura uma solução satisfatória entre o bom ajuste e o princípio da parcimônia (AKAIKE, 1974). O método AIC considera que os modelos apresentam melhor desempenho quanto mais simples (menor valor de AIC) for o modelo, portanto, o método impõe uma penalidade à complexidade. Como o logaritmo da função de verossimilhança $L(\beta)$ cresce com o aumento do número de parâmetros do modelo, uma proposta razoável seria encontrarmos o modelo com menor valor para a função

$$AIC = -L(\hat{\beta}) + p \tag{2.12}$$

em que p denota o número de parâmetros. No caso do modelo normal linear podemos mostrar que AIC fica expresso, quando σ^2 é desconhecido, na forma

$$AIC = nlog\left\{D(y; \hat{\mu}/n)\right\} + 2p, \tag{2.13}$$

em que $D(y; \hat{\mu}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$.

2.6.5 Critério de Informação Bayesiano - BIC

O Critério de informação Bayesiano (BIC), também chamado de Critério de Schwarz, foi proposto por Schwarz (SCHWARZ et al., 1978), e é um critério de avaliação de modelos definido em termos da probabilidade a posteriori, sendo assim chamado porque Schwarz deu um argumento Bayesiano para prová-lo. É dado por:

$$BIC = -2\,log f(x_n|\theta) + p\,log\,n, \tag{2.14}$$

em que $f(x_n|\theta)$ é o modelo escolhido, p é o número de parâmetros a serem estimados e n é o número de observações da amostra ($x_n = (x_1, x_2, ..., x_n)$).

2.6.6 Considerações AIC e BIC

Vale apenas salientar algumas características dos critérios AIC e BIC (EMILIANO, 2009) :

• Tanto o AIC quanto o BIC fundamentam-se na verossimilhança, impondo entretanto diferentes penalizações;

• O AIC e o BIC servem para comparar modelos encaixados, mas podem ser aplicados também em modelos não encaixados;

• Para n > 8, o valor do AIC para um determinado modelo será sempre menor que o valor do BIC;

• O AIC e o BIC servem para comparar quaisquer quantidade de modelos, e não somente dois;

• O AIC e o BIC são critérios assintóticos e já existem correções para estes;

• O AIC e o BIC servem para estudar estruturas de covariâncias;

• Se somente modelos ruins forem selecionados, o AIC fará a seleção do melhor dentre eles.

2.6.7 Deviance

A análise de deviance é uma generalização da análise da variância para os modelos lineares generalizados, visando obter, a partir de uma sequência de modelos, cada um incluindo mais termos do que os anteriores, os efeitos de fatores, covariáveis e suas interações. Dada uma sequência de modelos encaixados, utiliza-se a deviance como uma medida de discrepância do modelo e forma-se uma tabela de diferença de deviances.

Seja $M_{p_1}, M_{p_2}, ..., M_{p_r}$ uma sequência de modelos encaixados de dimensões respectivas $p_1 < p_2 < ... < p_r$, matrizes dos modelos $X_{p_1}, X_{p_2}, ..., X_{p_r}$ e deviances $D_1 >$

$D_2 > ... > D_r$, tendo os modelos a mesma distribuição e a mesma função de ligação. Estas desigualdades entre as deviances, em geral, não se verificam para a estatística de Pearson X^2 generalizada e, por esta razão, a comparação de modelos encaixados é feita, principalmente, via função deviance (DEMÉTRIO, 2001).

Sejam os modelos M_p e M_q $(p < q)$ com p e q parâmetros, respectivamente. A estatística $D_p - D_q$ com $(q - p)$ graus de liberdade é interpretada como uma medida de variação dos dados, explicada pelos termos que estão em M_q e não estão em M_p, incluídos os efeitos dos termos em M_p e ignorando quaisquer efeitos dos termos que não estão em M_q . Tem-se, assintoticamente, para ϕ conhecido, que

$$S_p - S_q = \frac{1}{\phi}\left(D_p - D_q\right) \sim X^2_{q-p}, \tag{2.15}$$

que é simplesmente o teste da razão de verossimilhanças. Se ϕ é desconhecido, deve-se obter uma estimativa $\hat{\phi}$ consistente, de preferência baseada no modelo maximal (com m parâmetros), e inferência pode ser baseada na estatística F, dada por

$$F = \frac{\left(D_p - D_q\right)/\left(q - p\right)}{\hat{\phi}} \sim F_{q-p,n-m} \tag{2.16}$$

Para a distribuição normal, temos:

$$F = \frac{\left(SQRes_p - SQRes_q\right)/\left(q - p\right)}{SQRes_m/\left(n - m\right)} \sim F_{q-p,n-m} \;\; (exata)\,. \tag{2.17}$$

3 Materiais e Métodos

Os dados a serem trabalhados foram coletados e produzidas a partir do questionário de avaliação médica do paciente do Programa Multidisciplinar de Tratamento do Tabagismo. Esse questionário é dividido nas seguintes seções: informações pessoais, identificação, tabagismo, histórico pessoal, histórico pessoal feminino, histórico familiar, uso de medicações, exame físico e anamnese. Das 117 variáveis de 714 usuários do Programa Multidisciplinar de Tratamento do Tabagismo dentre os anos de 2013-2017, selecionou-se 31 variáveis relevantes para este estudo, variáveis que serão descritas no apêndice A.

Para alcançar os objetivos que foram propostos nesse trabalho, ao recebermos os dados vindos dos questionários, fizemos a limpeza dos dados a partir do programa `Open Refine` para a retirada de dados inconsistentes e também organizá-los de uma maneira que possamos compreender o que aqueles dados representam ao nosso estudo. Cada pergunta foi analisada a fim de compreender os fatores que as respostas estavam inseridas. Através dessa análise inserimos uma nova coluna, denominada *CES*, com status binário 0 e 1, onde 0 indica o paciente que não concluiu o tratamento e 1 indica o paciente que concluiu o tratamento, ou seja, se houve ou não a cessação do tabagismo. O paciente cessa o tabagismo quando para de utilizar o fumo, ou seja, segundo o acompanhamento que foi feito através das fichas conseguimos identificar quando esse paciente cessou através das respostas à variável alteração de consumo que tem 6 etapas, explicando de uma maneira mais prática, a cada retorno do paciente ao programa verificava-se o uso do tabaco entre quantos fumava e quantos passou a fumar durante aquele determinado período. Quando o paciente chegava em uso 0 significa dizer que ele cessou o tabagismo e se o mesmo ainda faz uso ou abandonou o tratamento classificamos como paciente que não cessou o uso do tabagismo.

Antes da escolha das variáveis, faremos uma análise descritiva dos dados onde apontaremos as principais informações coletadas a partir das respostas que obtivemos através do questionário. Feita a escolha das 31 variáveis que são de característica determinada em fatores, ou seja, sua resposta é de forma qualitativa. Com o auxílio do programa `R` com o pacote *"arules"* faremos uso da Algoritmo Apriori para vermos as relações existentes entre as variáveis, observando as suas devidas associações.

Após a aplicação do Algoritmo Apriori e definido as associações e logo após a poda, teremos um novo número de variáveis para que assim possamos prosseguir com o estudo de tais. Temos então que a partir disso, aplicaremos um modelo de Regressão Logística com as variáveis e logo após aplicaremos um método de seleção do melhor modelo, aplicaremos então o método Stepwise que já foi descrito por nós em sessões anteriores.

Lembrando que na regressão logística os erros seguem distribuição binomial e a significância é assegurada via Teste da Razão de Verossimilhança. Assim, para cada passo do procedimento a variável mais importante em termos estatísticos, é aquela que produz a maior mudança no logaritmo da verossimilhança em relação ao modelo que não contém a variável.

Todos esses passos e resultados serão descritos em nosso próximo capítulo.

4 Resultados e Discussão

Neste capitulo faremos uma breve análise descritiva das variáveis, exibir as associações feitas a partir da utilização do Algoritmo Apriori, o modelo de Regressão Logística obtido a partir das associações e o melhor modelo que foi escolhido a partir da utilização do Stepwise.

4.1 Análise Descritiva

Dos 714 questionários que foram respondidos, em relação ao sexo dos pacientes conforme a Tabela 1, 469 são do sexo feminino e 245 do sexo masculino. A partir disso, calculamos quantos deles no geral cessaram o tabagismo:

Tabela 1 – Número de pacientes atendidos pelo Programa que cessaram ou não o uso do Tabagismo

	0 - Não cessou	1 - Cessou	Total
Sexo	n (%)	n (%)	N (%)
Feminino	282 (39,5)	187 (26,2)	469 (65,7)
Masculino	155 (21,7)	90 (12,6)	245 (34,3)
Total	437 (61,2)	277 (38,8)	714 (100,0)

Aplicando o teste Quiquadrado para testar a associação entre as variáveis de cessação e sexo, com α=0,05 e seu p-valor=$0,4618$ rejeitamos a hipótese que as variáveis estão associadas entre sí, ou seja, o fator de ser do sexo feminino ou masculino não interfere na cessação do tabagismo. Como podemos observar nos dados descritos temos que o número de pessoas que não cessaram o uso do Tabaco é maior do que aqueles que conseguiram cessar tal uso em ambos os sexos.

Em relação as idades dos pacientes atendidos pelo programa; a média é de 53 anos de idade; com os números variando entre 21 e 88. Os registros foram alocados em 9 faixas etárias conforme é possível se observar na Tabela 2. A classe de mulheres entre 53 e 61 obteve a maior frequência foram 154 pacientes do sexo feminino, representando 21,8% do total de mulheres participantes.

Os pacientes do sexo masculino a média das idades também é de 53 anos, com os números variando entre 17 e 92. Os registros foram alocados em 9 faixas etárias conforme é observado na Tabela 2. A classe de homens entre 53 e 61 anos obteve a maior frequência, foram 55 pacientes, representando 7,8% do total de homens participantes.

As classes que obtiveram o maior número de pacientes que cessaram o uso do tabaco segundo a Tabela 2 entre o sexo feminino foram de 53-61 com 9,6% seguida da

classe que vai das idades de 44-52 anos com 7,2%, dentre as que cessaram com a sua menor frequência foi a classe que vai dos 80-88 anos com 0,4% seguido da classe entre os anos de 17-25 com o mesmo valor de 0,4%. Falando em relação as pacientes que não cessaram o uso do tabagismo, a classe com maior frequência é a de 53-61 com 12,2% seguida da classe das idades entre 44-52 com 10,0%. Dentre as que não cessaram com menor frequência está a faixa etária entre 17-25 com 0,6% e também a faixa etária de 80-88 com 0,3%.

Nos pacientes do sexo masculino, tem-se da Tabela 2 que as classes com maiores frequências entre os que cessaram o uso do tabaco é de 53-61 com 3,5% seguida da classe com as idades entre 62-70 com 2,6%, os que tiveram menor frequência foi as de 89-97 que não teve nenhum, ou seja, 0% seguida das classes com as faixas etárias entre 17-25 e 80-88; com 0,3%. Em relação aos que não cessaram o uso do tabaco a classe com as maiores frequências são as de 35-43 e 53-61 com 4,5% e 4,2% e as de menores frequência são as classes de 80-88 com 0,4% e também de 89-97 com 0,1%.

Ressaltamos que o total da Tabela 2 é menor do que o total de pacientes que temos devido às respostas aos questionários pois, 5 pacientes do sexo feminino não forneceram sua data de nascimento e por isso o total da tabela referente ao sexo feminino é de 464, como também 2 pacientes do sexo masculino deixaram de fornecer sua data de nascimento, por esse motivo trabalhamos com o total de 243 referente ao sexo masculino. No total temos 707 participantes. Em relação as suas idades serem referidas ao ano de 2018, como citei em situações anteriores esses dados foram coletados durante os anos de 2013 até 2017 e suas idades não são referidas no questionário e sim suas datas de nascimento, assim analisamos os anos e tabulamos as idades.

Tabela 2 – Intervalos de classes das idades dos pacientes atendidos pelo Programa referido ao ano de 2018

	Não cessou	Cessou	Total
Faixa etária	n (%)	n (%)	N (%)
Feminino			
17-25	4 (0,6%)	3 (0,4%)	7 (1,0%)
26-34	20 (2,8%)	7 (1,0%)	27 (3,8%)
35-43	38 (5,4%)	18 (2,6%)	56 (8,0%)
44-52	71 (10,0%)	51 (7,2%)	122 (17,2%)
53-61	86 (12,2%)	68 (9,6%)	154 (21,8%)
62-70	43 (6,1%)	29 (4,1%)	72 (10,2%)
71-79	15 (2,1%)	6 (0,8%)	21 (2,9%)
80-88	2 (0,3%)	3 (0,4%)	5 (0,7%)
Masculino			
17-25	6 (0,8%)	2 (0,3%)	8 (1,1%)
26-34	13 (1,8%)	7 (1,0%)	20 (2,8%)
35-43	32 (4,5%)	16 (2,3%)	48 (6,8%)
44-52	26 (3,7%)	11 (1,6%)	37 (5,3%)
53-61	30 (4,2%)	25 (3,5%)	55 (7,7%)
62-70	27 (3,9%)	18 (2,6%)	45 (6,5%)
71-79	16 (2,3%)	8 (1,1%)	24 (3,4%)
80-88	3 (0,4%)	2 (0,3%)	5 (0,7%)
89-97	1 (0,1%)	0 (0%)	1 (0,1%)
Total	433 (61,2%)	274 (38,8%)	707 (100%)

Todas as informações contidas na Tabela 2 podem ser expressas através do gráfico de barras que temos abaixo. Confirmando de modo visual aquilo que a tabela nos trás

através dos números. Como é perceptível, o número de pacientes do sexo feminino é maior do que a do sexo masculino, quem não cessou o tabagismo é maior do que aqueles que cessaram devido ao atendimento do programa, além de percebermos as classes de idades com maiores e menores frequências.

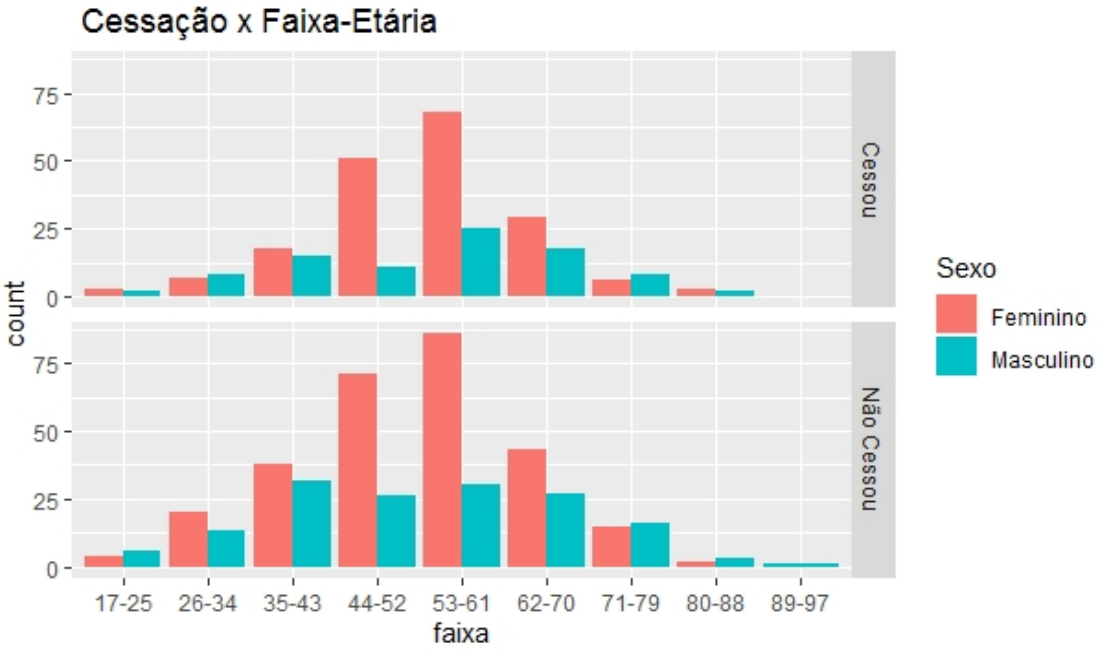

Figura 6 – Gráfico de Barras

Na Tabela 3 observa-se a frequência dos pacientes atendidos pelo programa que são casados; 316 os quais são classificados entre os que cessaram ou não o tabagismo: não cessou - 177 (56%) e cessou - 139 (44%), que por muita das vezes procuram o tratamento pela questão familiar. Ainda falando sobre o estado civil dos pacientes, a grande procura também por parte dos solteiros com 214 pacientes dividindo-se entre os que cessaram e os que não cessaram: 65 (30,4%) e 149 (69,9%). Tem-se indícios nesses dados é que a diferença entre os que cessaram e não cessaram entre os casados foi menor do que os solteiros , ou seja, podemos dizer que as pessoas casadas levam o tratamento mais a sério do que os solteiros como pode-se observar pelas porcentagens obtidas.

Tabela 3 – Número de pacientes atendidos pelo Programa segundo o Estado civil

	Não cessou	Cessou	Total
	n (%)	n (%)	N (%)
Estado Civil			
Solteiro (a)	149 (21,0%)	65 (9,2%)	214 (30,2%)
Casado (a)	177 (25,0%)	139 (19,6%)	316 (44,6%)
Divorciado (a)	62 (8,8%)	48 (6,8%)	110 (15,6%)
Viúvo (a)	45 (6,3%)	23 (3,3%)	68 (9,6%)
Total	433 (61,1%)	275 (38,9%)	708 (100%)

Em relação a escolaridade como temos na Tabela 4, a maioria possui de 9 a 12 anos 168 (23,6%), seguido pelos pacientes que possui o entre 6 e 8 anos 138 (19,4%) e que possui de 3 a 5 anos 112 (15,8%), e temos uma boa participação dos pacientes que possuem mais de 16 anos de escolaridade.

Tabela 4 – Número de pacientes atendidos pelo Programa segundo a Escolaridade

	Não cessou	Cessou	Total
	n (%)	n (%)	N (%)
Escolaridade			
Analfabeto	32 (4,5%)	16 (2,2%)	48 (6,7%)
Até 3 anos	27 (3,8%)	14 (2,0%)	41 (5,8%)
3 a 5 anos	67 (9,4%)	45 (6,3%)	112 (15,7%)
6 a 8 anos	87 (12,2%)	51 (7,2%)	138 (19,4%)
9 a 12 anos	95 (13,4%)	73 (10,2%)	168 (23,6%)
13 a 15 anos	69 (9,7%)	41 (5,8%)	110 (15,5%)
Mais de 16 anos	58 (8,2%)	36 (5,1%)	94 (13,3%)
Total	435 (61,2%)	276 (38,8%)	711 (100%)

Observa-se segundo a Tabela 5 a raça dos pacientes onde a maioria dos pacientes que não cessaram o tabagismo e se denominam Pardos, representa 34,8% seguido dos que se denominam brancos, com 19,6%. Entre os que cessaram o tabagismo do total, temos que os que se denominam pardos tem a maior frequência com 22,2% seguido dos que se denominam brancos com 10,5%.

Tabela 5 – Número de pacientes atendidos pelo Programa segundo a Raça

	Não cessou	Cessou	Total
	n (%)	n (%)	N (%)
Raça			
Amarela	5 (0,7%)	4 (0,6%)	9 (1,3%)
Branca	139 (19,6%)	74 (10,5%)	213 (30,1%)
Indígena	5 (0,7%)	2 (0,3%)	7 (1,0%)
Pardo	246 (34,8%)	157 (22,2%)	403 (57,0%)
Preta	37 (5,2%)	38 (5,4%)	75 (10,6%)
Total	432 (61,0%)	275 (39,0%)	707 (100%)

Observando a Tabela 6 com os dados referente a religião, temos que 498 (70%) dos pacientes se declaram em sua religião católica, a diferença entre os que não cessaram e cessaram é de 286 (40,2%) e 212 (29,8%). Logo após os pacientes que se declaram evangélicos 108 (15,1%), a diferença entre os que não cessaram e cessaram foi de modo 75 (10,5%) e 33 (4,6%). Dois pacientes não informaram sobre sua religião. Ao perguntar se são praticantes da religião, dos que se declaram praticantes representam 61,7% dos pacientes. Entre os que não cessaram ou cessaram temos 242 (35,9%) e 174 (25,8%), ou seja, aqueles que se dizem praticantes e cessaram o tabagismo ainda são minoria entre os pacientes.

Tabela 6 – Religião citada pelos pacientes e resposta sobre a prática da religião.

	Não Cessou	Cessou	Total
	n (%)	n (%)	N (%)
Religião			
Católica	286 (40, 2%)	212 (29, 8%)	498 (70, 0%)
Espírita	11 (1, 6%)	6 (0, 8%)	17 (2, 4%)
Evangélica	75 (10, 5%)	33 (4, 6%)	108 (15, 1%)
Não possui	54 (7, 6%)	24 (3, 4%)	78 (11, 0%)
Outra	10 (1, 4%)	1 (0, 1%)	11 (1, 5%)
Total	436 (61, 3%)	276 (38, 7%)	712 (100%)
Praticantes			
Sim	242 (35, 9%)	174 (25, 8%)	416 (61, 7%)
Não	168 (24, 8%)	91 (13, 5%)	259 (38, 3%)
Total	410 (60, 7%)	265 (39, 3%)	675 (100%)

4.2 Escolha do Modelo

O vetor reposta Y é binário, ou seja, assume o valor 0 quando o paciente não cessou o tabagismo e 1 quando paciente cessou o tabagismo. As variáveis explicativas que foram selecionadas inicialmente são: sexo, religião, estado civil, praticante da religião, cor, atividade profissional, após palestra (essa variável significa que se após assistir a palestra o paciente parou,diminuiu, aumentou ou continuou sem alteração o uso do tabagismo), aconselhamento médico sobre parar de fumar, reside com fumante, fator de influência, cigarro utilizado, aviso médico sobre hipertensão, uso de drogas, parente diabético, portador de artrite, diagnóstico de gastrite, diagnóstico de tireoide, parente hipertenso, parente obeso, parente asmático, parente cardiopata, escolaridade, se já fez dieta, renda familiar mensal, quantidade de pessoas que vivem com a renda, renda individual mensal, consumo de bebidas alcoólicas, prática de exercícios físicos, consumo de produtos dietéticos, usa medicação, diagnóstico de diabetes. Ou seja, temos 31 variáveis explicativas para modelar nossa resposta. Com o auxilio do programa `R`, com o pacote *arules* utilizamos as técnicas do algoritmo Apriori, ou seja, regras para que pudesse trabalhar com aquelas que mais explicassem o cessar ou não do tabagismo.

Na Tabela 7 temos como exemplo algumas das regras obtidas a partir da execução dos dados com o algoritmo Apriori. Como são muitas informações diversas regras foram criadas e assim ficaria muito extenso colocá-las aqui por completo, assim, exemplos de algumas regras são expostas abaixo:

Tabela 7 – Algumas regras obtidas com o algoritmo Apriori

Parentesco Hipertensão = Parentes de 1º e 2º grau	⟶ Cessação = 0
Parentesco Cardiopata = Parentes de 1º e 2º grau	⟶ Cessação = 0
Parentesco Diabético = Parentes de 1º e 2º grau	⟶ Cessação = 0
Parentesco Obesidade = Não há	⟶ Cessação = 0
Parentesco Hipertensão = Pai	⟶ Cessação = 1
Parentesco Asmático = Não há	⟶ Cessação = 1
Dieta = Não	⟶ Cessação = 1
Escolaridade = 9 a 12 anos	⟶ Cessação = 1
Parentesco Obesidade = Avós, Tios ou Primos	⟶ Cessação = 1

Após a execução dos dados com as regras e utilizando a poda (nessa etapa o algoritmo deve determinar se todos os subconjuntos são frequentes,se pelo menos um deles for infrequente então ele é podado, ou seja, reduzido), diminuímos as variáveis explicativas, assim como foi dado no exemplo na página 15 do algoritmo apriori. De 31 variáveis explicativas, 19 variáveis foram selecionadas pelo algoritmo Apriori. Para o ajuste da cessação do tabagismo utilizamos o procedimento *stepwise* e verificamos qual o melhor modelo que explica os dados. Após feito o *stepwise* temos então 16 ajustes, sendo o primeiro ajuste com o modelo completo e os demais seguindo as etapas do processo, a descrição dos modelos obtidos após o *stepwise* se encontra no Apêndice B desse trabalho e a descrição das variáveis que foram utilizadas para os ajustes se encontra na Tabela 8. Vejamos então as informações aqui descritas :

Tabela 8 – Descrição das variáveis após o algoritmo Apriori para a regressão logística

Variável	Identificação
x_1	aconselhado parar
x_2	estado civil
x_3	mora com fumante
x_4	fator de influência
x_5	aviso médico sobre hipertensão
x_6	portador de artrite
x_7	diagnóstico de diabete
x_8	fez alguma dieta
x_9	escolaridade
x_{10}	parente diabético
x_{11}	renda familiar mensal
x_{12}	prática de exercícios
x_{13}	religião
x_{14}	diagnóstico de gastrite
x_{15}	uso de drogas
x_{16}	bebida alcoólica
x_{17}	quantidade de pessoas que vivem com a renda
x_{18}	parente cardiopata
x_{19}	parente obeso

A partir dos modelos adquiridos pelo procedimento *stepwise* contido no Apêndice B desse trabalho, analisando os valores do AIC e BIC, fazendo as comparações e análises

dos ajustes obtidos temos que o melhor modelo que representa os dados aqui estudado é o ajuste 16 que contém as seguintes variáveis: Estado Civil, Mora com fumante,Uso de drogas e Religião , ou seja, o modelo apresenta menor AIC e BIC. Com isso, seguiremos apresentado resultados com o modelo ajustado escolhido. Para uma melhor visualização da tabela colocamos os nomes dos ajustes com uma pequena sigla, por exemplo, ajuste 1 será A1 e assim por diante até o último ajuste, essas informações constam na Tabela 9 a seguir:

Tabela 9 – AIC e BIC

Modelo	AIC	BIC
A1 (completo)	845,1287	1159,638
A2	834,2092	1118,141
A3	825,1145	1082,838
A4	815,7792	1042,925
A5	808,1240	1004,692
A6	801,0540	980,1497
A7	794,5322	947,4187
A8	788,4106	910,7198
A9	783,3787	888,2152
A10	781,3913	881,8596
A11	779,4060	875,5061
A12	777,4309	869,1628
A13	775,5152	862,879
A14	773,8530	835,0076
A15	773,0497	829,8361
A16	772,1262	820,1763

A Tabela 10 mostra as estatísticas para o modelo ajustado após o *stepwise* que nos indicou o ajuste 16 que contém as seguintes variáveis: Estado civil, Mora com fumante, Uso de drogas e Religião.

Tabela 10 – Estatísticas do ajuste 16

	Estimativa	Erro padrão	Estatística z	Pr(Z>|z|)
Intercepto	-0,9457	0,6719	-1,407	0,15929
Estado Civil				
Casado (a)				
Divorciado (*a*)	-0,2368	0,2541	-0,932	0,35141
Solteiro (*a*)	-0,5984	0,2101	-2,849	0,00439 **
Viúvo (*a*)	-0,1577	0,3209	-0,491	0,62315
Mora com fumante				
Sim	-0,3520	0,1944	-1,811	0,07013 .
Uso de drogas				
É usuário				
Ex- usuário	0,3885	0,7405	0,525	0,59985
Nunca usou	1,0095	0,6642	1,520	0,12858
Religião				
Católica				b
Espírita	-0,0436	0,5455	-0,080	0,93629
Evangélica	-0,5533	0,2541	-2,177	0,02947 *
Não possui	-0,1705	0,2941	-0,580	0,56202
Outra	-1,8213	1,0779	-1,690	0,09111 .

*** Significante a 0,001 ** Significante a 0,01 * Significante a 0,05 . Significante a 0,10

Observando os valores que esse ajuste nos retornou, observa-se que a variável *Uso de drogas* não foi significativo podemos então retirar essa variável do modelo e observar como as demais variáveis se comportam. Além disso, fizemos o agrupamento das variáveis que restam no modelo, por exemplo, *Religião* fixamos sua base significativa que de acordo com a Tabela 10 é 'Evangélica' e uni-se as demais denominando-as de "Outros"que são as demais religiões citadas, do mesmo modo acontece com o *Estado Civil* . Na Tabela 11 estão expostos as estatísticas do modelo que foi ajustado com a retirada da variável *Uso de drogas* e o agrupamento das demais e chamaremos de ajuste 17. Observando-se os valores das estatísticas e o seu P valor, temos que a variável *Mora com fumante* não foi significativa, então faremos um novo modelo retirando também essa variável para observar os valores do modelo proposto.

Tabela 11 – Estatísticas do ajuste 17

	Estimativa	Erro padrão	Estatística z	Pr(Z>|z|)
Intercepto	-0,1327	0,1192	-1,114	0,26543
Estado Civil				
Solteiro (a)	-0,5442	0,1938	-2,808	0,00499 **
Mora com fumante				
Sim	-0,3085	0,1908	-1,617	0,10587
Religião				
Evangélica	-0,5099	0,2490	-2,048	0,04054 *

*** Significante a 0,001 ** Significante a 0,01 * Significante a 0,05 . Significante a 0,10

Na Tabela 12 temos os valores das estatísticas de um novo ajuste com a retirada da variável *Uso de drogas* e *Mora com fumante*, nesse modelo ajustado todas as variáveis são significativas e por final escolhemos esse ajuste como modelo final, ou seja, o modelo escolhido para a representação dos dados aqui estudado é o que contém as variáveis: *Estado Civil* e *Religião*. O modelo é significativo ao nível de 5%.

Tabela 12 – Estatísticas do ajuste 18 - ajuste final

	Estimativa	Erro padrão	Estatística z	Pr(Z>|z|)
Intercepto	-0,2175	0,1072	-2,030	0,04239 *
Estado Civil				
Solteiro (a)	-0,5638	0,1931	-2,919	0,00351 **
Religião				
Evangélica	-0,5155	0,2484	-2,075	0,03795 *

*** Significante a 0,001 ** Significante a 0,01 * Significante a 0,05 . Significante a 0,10

Comparamos ainda os valores dos dois ajustes em relação a Deviance, AIC e BIC, os valores estão dispostos na Tabela 13 e podemos perceber que o ajuste com agrupamento e sem as variáveis *Uso de drogas* e *Mora com fumante* é a melhor de ser utilizada, ou seja, o ajuste 18 é que pode representar melhor os dados de um modo mais significativo, seguindo o critério do BIC que é menor, apenas o AIC que deu uma diferença de $0,65$ do

ajuste 17, porém, é bem pequena a diferença. Com isso, escolhemos o ajuste 18 pois ele é mais simples e representará bem os dados aqui estudado.

Tabela 13 – Comparação dos ajustes em relação ao AIC e BIC

	AIC	BIC
Ajuste 16	772,1262	820,1763
Ajuste 17	770,8582	788,3309
Ajuste 18	771,5068	784,6113

Na Tabela 14 temos a razão de chance do modelo escolhido. Fazendo a leitura da segunda coluna a chance de um paciente do programa cessar o tabagismo se reduz em 43% quando ele é solteiro comparado com os outros estados civis. O intervalo de confiança de 95% da verdadeira razão de chances entre pacientes solteiros com relação aos outros estados civis e fixada a religião pode estar entre 0,39 e 0,83. Ao comparar a religiosidade observa-se que a chance de um paciente do programa se reduz em 40% quando é evangélico comparado com outras religiões citadas no questionário e fixado o estado civil; o intervalo de confiança de 95% da verdadeira razão de chances entre os pacientes evangélicos em relação as demais religiões é de 0,36 e 0,96.

Tabela 14 – Razão de chances do modelo

	RC	IC (95%)	Pr(Z>|z|)
Intercepto	0,80450	(0,65149;0,9920)	0,042391 *
Estado Civil			
Solteiro(a)	0,56903	(0,38763;0,8273)	0,003506 **
Religião			
Evangélica	0,59717	(0,36205;0,9619)	0,037953 *

Para se ter uma melhor visualização dos dados ajustados utiliza-se o gráfico de envelope que é construído a partir de intervalos de confiança simulados (PAULA, 2004). Observe na Figura 7 que todos os pontos têm resíduos ajustados dentro do envelope do modelo. Assim, o modelo escolhido, que foi o ajuste 18, é capaz de explicar bem a cessação do tabagismo. Portanto, a perda no AIC é justificada, pois o grau de explicação dos dados pelo ajuste 18 é muito superior ao ajuste 16. Pode-se então dizer que o ajuste 18 que contém as variáveis *Estado Civil* e *Religião* são fatores importantes na cessação do tabagismo. No que se refere a religiosidade, (GEHLING et al., 2011) apontam que apesar de ter menor prevalência de fumantes na população, evangélicos possuem risco de dependência aumentado em relação aos católicos. A partir destes resultados que foram aqui apresentados, pode-se levantar a hipótese de que a dificuldade de largar o cigarro pode ser um fator que influência positivamente a evasão do programa.

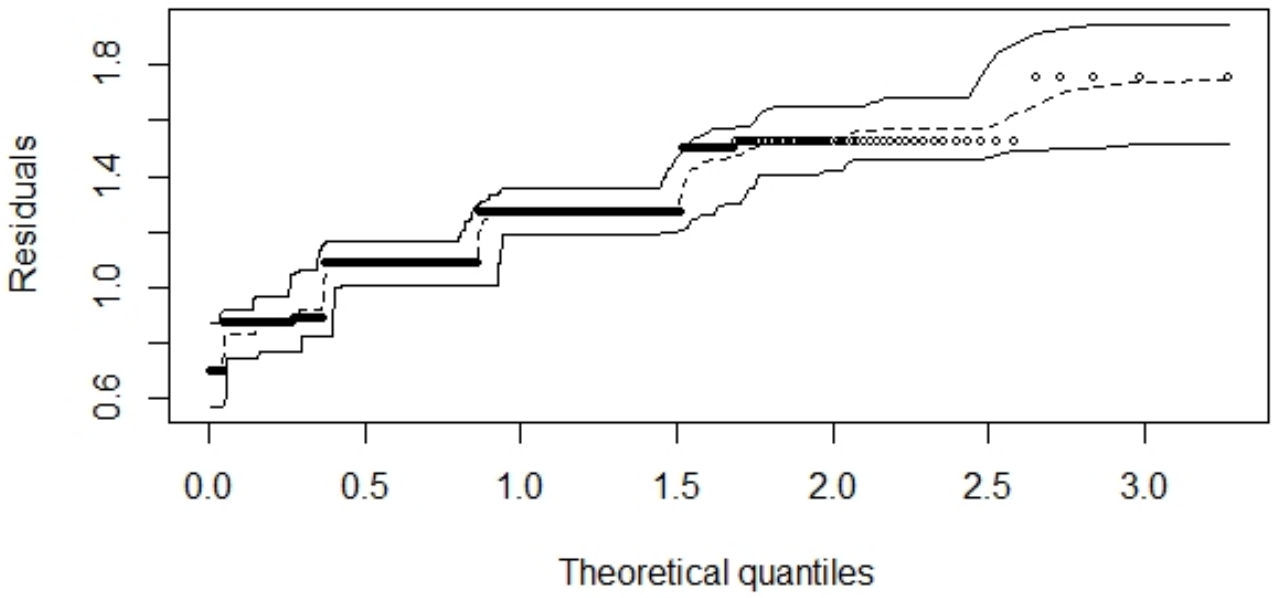

Figura 7 – Gráfico de envelope

5 Conclusão

A partir do que foi exposto neste trabalho, pode-se perceber que existem fatores que influenciam a cessação do tabagismo, tais seu estado civil e religião. Utilizou-se o programa de limpeza de dados *Open Refine* que nos auxiliou a compreender a natureza dos dados e como poderíamos ver as variáveis e assim poder classificá-las e trabalhar com os dados sem problemas posteriores. As Regras de Associação, com a aplicação do algoritmo Apriori são ferramentas úteis para ver a relação entre as variáveis e poder dar respostas a tantas perguntas que temos em relação aos dados trabalhados. De 31 variáveis que possivelmente poderia explicar sobre a cessação do tabagismo dos pacientes atendidos pelo programa, com as etapas descritas em nossos objetivos principal e secundário chegamos a um modelo que representa de uma melhor maneira os dados. Com o *stepwise* obteve-se 16 ajustes para as escolhas das variáveis para esse modelo, com as estatísticas que obtivemos com esse ajuste, pode-se utilizar de outras técnicas para uma melhor tomada de decisão, tal como agrupar as variáveis base e a retirada das que não fossem significativas. A escolha e comparação de tais ajustes foi realizada a partir do AIC e BIC, com aplicações claras do que foi estudado com esse trabalho. A partir do gráfico de envelope podemos observar como os dados estão bem ajustados ao modelo e pode explicar e justificar a escolha do ajuste 18 que continha apenas duas variáveis que são significativas, Estado Civil e Religião. Assim, pode-se dizer que os objetivos propostos nesse trabalho foram alcançados. Por mais que o número de pessoas que não cessaram o uso do tabagismo seja maior entre os pacientes que procuraram ajuda do programa; é importante que mais pesquisas sejam realizadas para que indicadores façam a diferença no modo de tratamento e que futuramente o número de pessoas que cessem o tabagismo seja maior do que foi apresentado nos dados estudados nesse trabalho.

Referências

AKAIKE, H. A new look at the statistical model identification. In: *Selected Papers of Hirotugu Akaike*. [S.l.]: Springer, 1974. p. 215–222. Citado na página 22.

CAVALCANTE, T. M. O controle do tabagismo no brasil: avanços e desafios. *Archives of Clinical Psychiatry*, v. 32, n. 5, p. 283–300, 2005. Citado na página 11.

CORRAR, L. et al. Análise multivariada para os cursos de administração, ciências contábeis e economia. 2011. Citado na página 19.

DANTAS, D. R. G. et al. Prevalência e risco de tabagismo entre estudantes do ensino médio em cidade do nordeste do brasil. *Portuguese Journal of Public Health*, Karger Publishers, v. 35, n. 1, p. 44–51, 2017. Citado na página 14.

DEMETRIO, C. Modelos lineares generalizados em experimentação agrícola. *1º. ed. Piracicaba, SP: ESALQ/USP*, 2002. Citado na página 19.

DEMÉTRIO, C. G. B. *Modelos lineares generalizados em experimentação agronômica*. [S.l.]: USP/ESALQ, 2001. Citado na página 24.

EMILIANO, P. C. Fundamentos e aplicações dos critérios de informação: Akaike e bayesiano. *Universidade Federal de Lavras*, 2009. Citado na página 23.

FREITAS, P. H. d. Regressão logística na modelagem da probabilidade de vitória em jogos de futebol americano. Universidade Federal de Uberlândia, 2019. Citado na página 12.

GEHLING, I. et al. Prevalência da dependência ao tabaco na população urbana e ribeirinha em coari (am), 2010. Florianópolis, SC, 2011. Citado na página 35.

INCA. 2017. Disponível em: <https://www.inca.gov.br/noticias/no-dia-mundial-sem-tabaco-pesquisa-revela-que-gastos-com-o-tabagismo-somam-quase-r-57>. Citado na página 11.

KIST, B. et al. Anuário brasileiro do fumo. *Editora Gazeta Santa. Cruz do Sul*, 2004. Citado na página 14.

MALIK, U. *Association Rule Mining via Apriori Algorithm in Python*. 2018. Disponível em: <https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/>. Citado na página 16.

MONTEIRO, C. A. et al. Population-based evidence of a strong decline in the prevalence of smokers in brazil (1989-2003). *Bulletin of the World Health Organization*, SciELO Public Health, v. 85, p. 527–534, 2007. Citado na página 14.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Citado na página 19.

ORGANIZATION, W. H.; CONTROL, R. for I. T. *WHO report on the global tobacco epidemic, 2008: the MPOWER package*. [S.l.]: World Health Organization, 2008. Citado na página 11.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004. Citado 4 vezes nas páginas 19, 20, 22 e 35.

POSSAMAI, F. B. Uso de regressão logística para um estudo da reincidência criminal no sistema penitenciário mediciense. 2018. Citado na página 19.

REGULASUS. *Resumos Clínicos - Tabagismo*. 2015. Disponível em: <https://www.ufrgs.br/telessauders/documentos/protocolos_resumos/pneumologia_resumo_tabagismo_TSRS_20160321.pdf>. Citado na página 14.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 23.

SCOGNAMIGLIO, e. a. H. 2016. Disponível em: <http://reporterunesp.jor.br/wp-content/uploads/2016/12/goog.jpg>. Citado na página 12.

TARGA, C. *Mineração eficiente de regras de associação através da indexação de conjuntos candidatos*. Tese (Doutorado) — Dissertação de Mestrado, Universidade Federal Fluminense, 2002. Citado na página 12.

ZHAO, Y. *R and data mining: Examples and case studies*. [S.l.]: Academic Press, 2012. Citado na página 16.

Apêndices

APÊNDICE A – Variáveis Qualitativas Nominais e Ordinais utilizadas

As variáveis para esse estudo devem ser de modo qualitativa pelo algoritmo que foi utilizado, por esse modo trabalhamos com essas que possuem essa característica. No questionário que os pacientes respondiam, todas essas variáveis estavam como uma pergunta e então diminuímos a maioria delas para ficar de melhor acesso e visualização.

- *Variáveis qualitativas nominais*:

Sexo

"Feminino" "Masculino"

Religião

"Católica" "Outra" "Espírita" "Evangélica" "Não possui"

Praticante da religião

"Sim" "Não"

Cor

"Amarela-Oriental" "Branca" "Indígena" "Pardo-mulato" "Preta"

Estado civil

"Solteiro(a)" "Casado(a)" "Viúvo (a)" "Divorciado(a)" "Separado (a)"

Atividade profissional

"Aposentado" "Autônomo" "Desempregado"

"Dona de casa" "Funcionário público" "Não se aplica"

"Outros" "Trabalhador c/ carteira assinada" "Voluntário"

Após palestra

"Aumentou" "Diminuiu" "Não houve alteração" "Parou de fumar"

Aconselhado parar

"Sim" "Não"

Mora com fumante

"Sim" "Não"

Fator de influência para inicio do tabagismo

"Curiosidade" "Exemplo da Mãe" "Exemplo do Pai" "amigos/colegas"

"outros familiares" "Outros" "Propaganda"

Tipo de cigarro

"Cachimbo" "Cigarros artesanais" "Cigarros Industrializados"

Aviso médico sobre Hipertensão

"Sim" "Não"

Uso de drogas

"É usuário" "Ex-usuário" "Nunca usou drogas"

Parente diabético

"Avós, Tios ou Primos" "Não Há" "Sem informação" "Irmão/Irmã"

"Filho" "Mais de um parente de primeiro grau" "Pai" "Mãe"

"Parentes de 1° e 2° grau"

Portador de artrite

"Não" "Sim"

Diagnóstico de gastrite

"Não" "Sim"

Diagnóstico de Tireóide

"Não" "Sim"

Parente hipertenso

"Avós, Tios ou Primos" "Não Há" "Sem informação" "Irmão/Irmã"

"Filho" "Mais de um parente de primeiro grau" "Pai" "Mãe"

"Parentes de 1° e 2° grau"

Parente obeso

"Avós, Tios ou Primos" "Não Há" "Sem informação" "Irmão/Irmã"

"Filho" "Mais de um parente de primeiro grau" "Pai" "Mãe"

"Parentes de 1° e 2° grau"

Parente asmático

"Avós, Tios ou Primos" "Não Há" "Sem informação" "Irmão/Irmã"

"Filho" "Mais de um parente de primeiro grau" "Pai" "Mãe"

"Parentes de 1° e 2° grau"

Parente cardiopata

"Avós, Tios ou Primos" "Não Há" "Sem informação" "Irmão/Irmã"

"Filho" "Mais de um parente de primeiro grau" "Pai" "Mãe"

"Parentes de 1° e 2° grau"

Dieta

"Não" "Sim"

Diagnóstico de diabético

"Não" "Sim, tipo I" "Sim, tipo II"

Consome produtos dietéticos?

"Não" "Sim"

- *Variáveis Qualitativas Ordinais*:

Escolaridade

"Analfabeto" "Até 3 anos" "3 a 5 anos" "6 a 8 anos"

"9 a 12 anos" "13 a 15 anos" "Mais de 16 anos"

Renda familiar mensal

"Menos de 500 reais" "501 a 750 reais" "751 a 1000 reais"

"1001 a 1500 reais" "1501 a 2500 reais" "2501 a 5000 reais"

"mais de 5001 reais"

Número pessoas que vive com renda familiar mensal

"1" "2" "3" "4" "5 ou mais"

Consumo de bebida alcoólica

"Nunca bebeu" "Não consome atualmente" "Até 5 doses por mês"

"De 5 a 10 por mês" "10 a 15 por mês" "Mais de 15 doses por mês"

"Diariamente"

Pratica Exercício

"Não" "Apenas do fim de semana" "2 a 3 vezes por semana"

"3 a 5 vezes por semana" "+5 vezes na semana"

Renda individual mensal

"Recebe benefício" "Menos de 500" "501 a 750 reais"

"750 a 1000 reais" "1001 a 1500 reais" '1501 a 2500 reais'

"2501 a 5000 reais" "Mais de 5001 reais" "Sem informação"

APÊNDICE B – Ajustes

Tabela 15 – Ajustes após o *stepwise*

A1	$\beta_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19}$
A2	$\beta_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19}$
A3	$\beta_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_{10} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{18} + x_{19}$
A4	$\beta_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_{10} + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{19}$
A5	$\beta_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_{12} + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{19}$
A6	$\beta_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{19}$
A7	$\beta_0 + x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_{13} + x_{14} + x_{15} + x_{16} + x_{17} + x_{19}$
A8	$\beta_0 + x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_{13} + x_{14} + x_{15} + x_{16} + x_{17}$
A9	$\beta_0 + x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_{13} + x_{14} + x_{15} + x_{16}$
A10	$\beta_0 + x_1 + x_2 + x_3 + x_6 + x_7 + x_8 + x_{13} + x_{14} + x_{15} + x_{16}$
A11	$\beta_0 + x_2 + x_3 + x_6 + x_7 + x_8 + x_{13} + x_{14} + x_{15} + x_{16}$
A12	$\beta_0 + x_2 + x_3 + x_6 + x_7 + x_8 + x_{13} + x_{15} + x_{16}$
A13	$\beta_0 + x_2 + x_3 + x_7 + x_8 + x_{13} + x_{15} + x_{16}$
A14	$\beta_0 + x_2 + x_3 + x_7 + x_8 + x_{13} + x_{15}$
A15	$\beta_0 + x_2 + x_3 + x_7 + x_{13} + x_{15}$
A16	$\beta_0 + x_2 + x_3 + x_{13} + x_{15}$