Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

USO DE CIÊNCIA DE DADOS PARA ESTUDO DAS VARIÁVEIS ASSOCIADAS À DEPRESSÃO EM FUMANTES

Felipe Rafael de Souza

JUIZ DE FORA
JULHO, 2019

USO DE CIÊNCIA DE DADOS PARA ESTUDO DAS VARIÁVEIS ASSOCIADAS À DEPRESSÃO EM FUMANTES

Felipe Rafael de Souza

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Heder Soares Bernardino

JUIZ DE FORA

JULHO, 2019

USO DE CIÊNCIA DE DADOS PARA ESTUDO DAS VARIÁVEIS ASSOCIADAS À DEPRESSÃO EM FUMANTES

Felipe Rafael de Souza

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Heder Soares Bernardino
Prof. Dr. Heder Soares Bernardino

Luciana Conceição Dias Campos
Prof. Dra. Luciana Conceição Dias Campos

Victor Ströele de Andrade Menezes
Prof. Dr. Victor Ströele de Andrade Menezes

JUIZ DE FORA

04 DE JULHO, 2019

Aos meus amigos e irmãos.
Aos pais, pelo apoio e sustento.
E, especialmente, à minha noiva, Gabriella,
pelo apoio em todo o tempo.

Resumo

Desde o século XVII, com a equação de Torricelli, o ser humano busca formas de prever o futuro através da modelagem de fenômenos. Saber como uma mudança influencia em um fenômeno, pode ser muito vantajoso. Por exemplo, descobrir a partir de um conjunto de dados, se a ação de uma empresa vai subir ou descer de valor pode trazer grande lucro. Ou, dado uma série de sintomas, prever uma doença. A taxa de sucesso na cessação do consumo de tabaco é influenciada pela presença de depressão. Logo, identificar pacientes depressivos e ajustar o tratamento para esses casos, pode gerar melhora no sucesso da intervenção. Pode-se utilizar diversos modelos para essa tarefa, como redes neurais, máquinas de vetor suporte, árvores de decisão ou programação genética. O Viva Sem Tabaco é um sistema de intervenção online que oferece ferramentas para auxiliar a cessação do consumo de tabaco. Esse trabalho tem o objetivo de gerar modelos de classificação, utilizando árvores de decisão para indicar a presença de depressão nos usuários do Viva Sem Tabaco. Com base nos modelos obtidos, foram encontradas relações entre o consumo de álcool, dependência do tabaco, idade e desemprego com a presença de depressão em fumantes.

Palavras-chave:, Árvore de decisão, inferência de modelos, Depressão, Ciência de Dados, Aprendizado de Máquina.

Abstract

Since the seventeenth century, with the equation of Torricelli, the human being looks for ways to predict the future through the modeling of phenomena. Knowing how a change influences a phenomenon can be very advantageous. For example, figuring out a set of data, whether a company's stock will go up or down value can bring big profit. Or, given a series of symptoms, predict a disease. The success rate in cessation of smoking is influenced by the presence of depression. Therefore, identifying depressive patients and adjusting treatment for these cases may lead to improvement in the success of the intervention. Several models can be used for this task, such as neural networks, support vector machines, decision trees or genetic programming. Viva Sem Tabaco is an online intervention system that offers tools to help stop smoking. This work has the objective of generating classification models, using decision trees to indicate the presence of depression in Viva Sem Tabaco users. Based on the models obtained, relationships were found between alcohol consumption, tobacco dependence, age and unemployment with the presence of depression in smokers.

Keywords: Decision Tree, Model Inference, Depression, Data Science, Machine Learning.

Agradecimentos

A todos os meus parentes, pelo encorajamento e apoio. Aos meus pais, pelo apoio constante e pelo carinho. À minha irmã. Ao meu filho de quatro patas, Zeus. À minha noiva, Gabriella, pelo carinho, pelo apoio sem medidas, pelas horas me aguentando falar sobre aprendizado de máquina, pelas ajudas na revisão, pelas várias vezes vendo friends no BK e tudo mais.

Aos professores do Departamento de Ciência da Computação, pelos seus ensinamentos e aos funcionários do curso que, durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional. Em especial ao meu orientador, Heder, pelo apoio e compreensão. Ao Henrique e à Nathália pelo apoio no desenvolvimento do trabalho. A todos os amigos e colegas que participaram desses anos de enriquecimento pessoal e de conhecimento.

"Não entre em Pânico"

Douglas Adams.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC Departamento de Ciência da Computação

UFJF Universidade Federal de Juiz de Fora

1 Introdução

1.1 Apresentação do tema

A inferência de modelos possui aplicações nas mais diferentes áreas, tais como na Engenharia (SOUZA et al., 2016), Biologia (Veiga et al., 2015), Saúde, Economia (Bollen et al., 2011) e Ciência da Computação. Seu objetivo é descobrir relações entre variáveis relacionadas a um fenômeno de interesse. Por exemplo, a partir de um conjunto de registros contendo informações temporais e da demanda por um certo produto, é possível criar um modelo que represente esta demanda ao longo do ano. Esse modelo pode ser usado, por exemplo, para a contratação de funcionários temporários, a fim de suprir altas de demandas em determinadas épocas. Portanto, a geração de um bom modelo para um problema pode ter diversas implicações, tais como a diminuição dos custos de produção, detecção de possíveis falhas (Murray et al., 2005) e desenvolvimento de novos produtos.

Em diversas ocasiões é importante, além de saber prever um fenômeno, entender a relação de causa atribuída a ele. Como por exemplo, além de saber se um paciente está ou não com depressão, saber quais fatores estão relacionados a essa doença podem auxiliar na tomada de decisão do profissional responsável pelo seu tratamento (Pinto Gomide et al., 2018).

1.2 Problema

O tabagismo é a dependência física e psicológica de nicotina, substância presente em produtos como cigarros e charutos, que constitui o princípio ativo do tabaco. O consumo regular de tabaco é responsável por diversas doenças. Dados da OMS (WHO, 2017) indicam que 9 em cada 10 mortes por câncer de pulmão são causadas pelo tabagismo. Os dados ainda indicam que morrem mais pessoas de doenças relacionadas ao tabagismo do que de AIDS, álcool, drogas ilegais, assassinatos, suicídios e acidentes automobilísticos juntos (WHO, 2017).

O Viva Sem Tabaco[1] (Gomide et al., 2016) é um sistema de intervenção online que oferece ferramentas para auxiliar a cessação do consumo de tabaco. O sistema possui diversas informações socioeconômicas e históricas do consumo de tabaco dos usuários, além de resposta a questionários psicométricos que indicam, por exemplo: depressão, dependência do consumo de álcool e tabaco. A identificação de depressão é de grande importância para a eficiência do tratamento adequado ao usuário (Gomide et al., 2017). Nesse contexto, gerar um modelo que relacione informações socioeconômicas e psicométricas para indicar depressão é de grande valor.

1.3 Justificativa

A depressão está diretamente associada a diminuição da taxa de cessação de consumo de tabaco (Covey et al., 1990). Logo, detectar pacientes depressivos e entender os motivos associados à ela é importante para melhorar a eficácia da intervenção. Nesse cenário, a árvore de decisão se destaca dos algoritmos ditos caixa preta, por ser capaz de gerar modelos em forma simbólica, permitindo a interpretabilidade das suas soluções pelo especialista de domínio, a fim de extrair e/ou validar conhecimento acerca do fenômeno estudado.

1.4 Objetivos

Esse trabalho tem como objetivo realizar uma análise na base de dados do Viva Sem Tabaco, a fim de gerar modelos preditivos, interpretáveis, e para estudar a relação dos dados socioeconômicos e psicométricos com a presença de depressão nos usuários.

[1] www.vivasemtabaco.com.br

2 Fundamentação Teórica

Nesse capítulo são apresentados os principais conceitos relacionados ao tema proposto, bem como o ferramental e as métricas utilizadas para o desenvolvimento desse trabalho. Os conceitos referentes à classificação são apresentados na secção 2.1, à árvores de decisão na secção 2.2, à validação cruzada na secção 2.3, às métricas utilizadas na secção 2.4, à correlação na secção 2.5 e finalmente a secção mostra o questionário psicométrico PHQ.

2.1 Problemas de classificação

Reconhecer padrões em imagens, identificar espécies de animais, diagnosticar um paciente como saudável ou doente, são todos problemas de classificação. A classificação de dados é um problema de aprendizado de máquina, que consiste em determinar o rótulo de um objeto com base em seus atributos (Aggarwal, 2014). Existem diversas técnicas capazes de realizar essa tarefa e essas técnicas podem ser divididas em dois grupos (Khan et al., 2012):

- **Caixa preta**: Gera modelos em que as tomadas de decisão são, em geral, não interpretáveis ou de difícil interpretação, como em redes neurais ou máquinas de vetor suporte;
- **Caixa branca**: Gera modelos em que as tomadas de decisão são de fácil interpretação para o analista, como árvores de decisão ou programação genética.

A classificação de dados pode ser descrita como um mapeamento de um conjunto de dados em valores discretos, chamados de categorias. O objeto capaz de classificar um conjunto de dados é chamado de classificador. O classificador se baseia em dados prévios para o treinamento, logo, a classificação de dados é dita uma tarefa de aprendizado supervisionado (Augusto, 2009).

2.2 Árvore de Decisão

Árvore de decisão é um modelo de aprendizado supervisionado, utilizado para classificação ou regressão (Breiman, 2017). É composta por uma estrutura encadeada de condicionais na forma de uma árvore. As árvores de decisão podem ser definidas como uma estrutura de dados recursivas (Monard and Baranauskas, 2003), onde:

- As folhas dessa árvore representam o valor da variável objetivo.
- Cada nó interno (nó de decisão) representa o teste de um atributo onde cada filho desse nó gera uma sub-árvore contendo um nó folha (nó que não possui filhos) ou um nó de decisão.

Modelos de árvore de decisão utilizam a estratégia de dividir para conquistar (Garcia, 2003), onde um problema complexo é dividido em sub-problemas mais simples de maneira recursiva.

A Figura 2.1 ilustra uma árvore de decisão para definir qual meio de transporte usar pra se deslocar. Para utilizar o modelo, basta começar pela raiz da árvore, seguindo cada nó de teste até que seja encontrada uma raiz (Bicicleta ou Carro).

A árvore de decisão também pode ser representada como um conjunto de regras, onde cada regra tem seu início na raiz da árvore e caminha até uma de suas folhas. O exemplo apresentado na Figura 2.1 também pode ser representado na forma de um algoritmo, como é apresentado no Algoritmo 1.

Como apenas uma única regra da árvore de decisão é disparada na classificação de uma amostra, é possível representar o modelo da árvore de decisão, escrevendo condicionais que se iniciam na raiz e vão até cada uma das folhas de forma independente. O exemplo do Algoritmo 1 pode ser representado pelas regras:

- **Se** está chovendo, **então** Carro
- **Se** não está chovendo **e** você está atrasado **e** é horário de pico, **então** Bicicleta
- **Se** não está chovendo **e** você está atrasado **e** não é horário de pico, **então** Carro
- **Se** não está chovendo **e** você não está atrasado, **então** Bicicleta

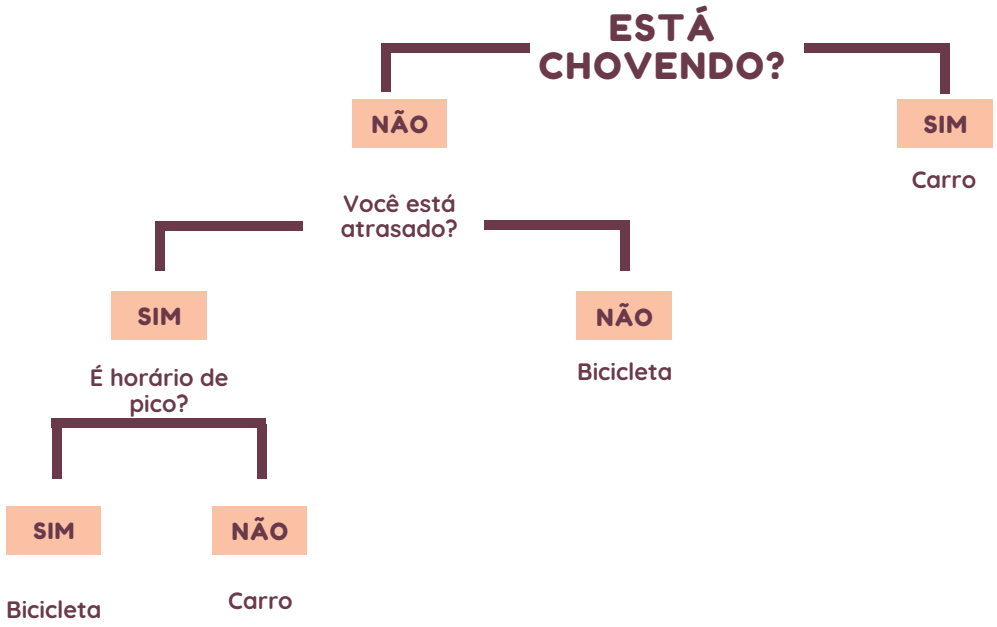

Figura 2.1: Árvore de decisão para definir meio de transporte.

Algoritmo 1: Representação algorítmica da árvore de decisão da Figura 2.1

```
1  se Está chovendo então
2  |  Carro
3  fim se
4  senão
5  |  se Você está atrasado então
6  |  |  se É horário de pico então
7  |  |  |  Bicicleta
8  |  |  fim se
9  |  |  senão
10 |  |  |  Carro
11 |  |  fim se
12 |  fim se
13 |  senão
14 |  |  Bicicleta
15 |  fim se
16 fim se
```

Existem diversos algoritmos de classificação que utilizam árvore de decisão, como por exemplo:

- C4.5 (Quinlan, 2014)
- ID3 (Quinlan, 1986)
- CART (Breiman et al., 1984)

Nesse trabalho será utilizado o algoritmo RPART (R Core Team, 2018), que é uma implementação de software livre do algoritmo proprietário de CART (Breiman et al., 1984). O RPART utiliza a partição recursiva binária para definir as regras de divisão dos nós. O processo é binário, pois cada nó dá origem a exatamente dois filhos; e recursivo, pois cada nó filho pode ser tratado como um nó pai, tendo como caso base da recursão encontrar uma folha (que representa a decisão do modelo).

2.3 Validação Cruzada

Algoritmos de aprendizado de máquina são, em geral, medidos com base no seu erro de previsão. Na maioria dos problemas do mundo real o erro não pode ser calculado de maneira determinística, sendo necessário uma estimativa do comportamento do modelo sobre novos dados. A validação cruzada (Efron, 1983) é uma estratégia para avaliação da capacidade de generalização de modelos e amplamente utilizada para prevenção de *overfitting*[2] (Hawkins, 2004). O cerne dessa técnica consiste na separação do conjunto de dados em subconjuntos disjuntos, onde parte desses subconjuntos são utilizados para realizar o treinamento do modelo, e outra parte para a validação (Efron and Tibshirani, 1997). Um método de validação cruzada é o *K-folds* (Kohavi et al., 1995).

O *K-folds* consiste em dividir o conjunto de dados em k subconjuntos, disjuntos entre si, e com tamanhos iguais (ou aproximadamente iguais, caso k não for divisível pelo número de registros) e utilizar k-1 subconjuntos para treinar o modelo e um subconjunto para testar, onde esse processo é repetido por k iterações, até que todos subconjuntos tenham sido usados como subconjuntos de teste. A Figura 2.2 ilustra a aplicação do

[2]Termo usado para descrever quando um modelo se ajusta muito bem ao conjunto de dados observado, porém é ineficaz para prever novos resultados

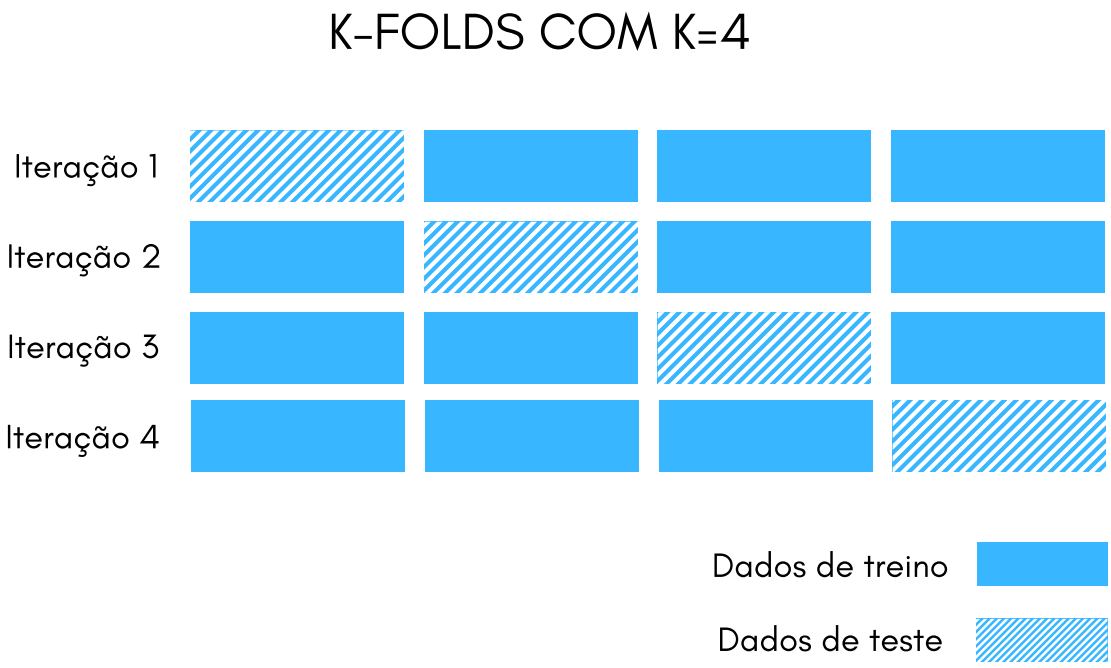

Figura 2.2: Ilustração da aplicação do *K-folds* (K=4) sobre uma base de dados.

K-folds (k=4) sobre um conjunto de dados. Após esse processo, é feita uma média das performances obtidas, sobre o conjunto de teste, de cada uma das K iterações, obtendo uma estimativa do erro de generalização do modelo. É importante destacar que não existe uma estimativa não enviesada para o erro do modelo sobre dados não vistos (Bengio and Grandvalet, 2004), mas a aproximação obtida utilizando a validação cruzada k-fold normalmente é uma aproximação suficientemente boa.

2.4 Métricas

2.4.1 Matriz de confusão

Matriz de confusão é um instrumento utilizado para problemas de classificação, que permite medir a performance de um modelo de predição (Stehman, 1997). A matriz de confusão consiste na representação em uma matriz de dimensão n×n, onde n é o número de classes do problema. As classes previstas pelo modelo são dispostas em uma das dimensões e as classificações reais dos dados na outra dimensão, ou seja, se as predições

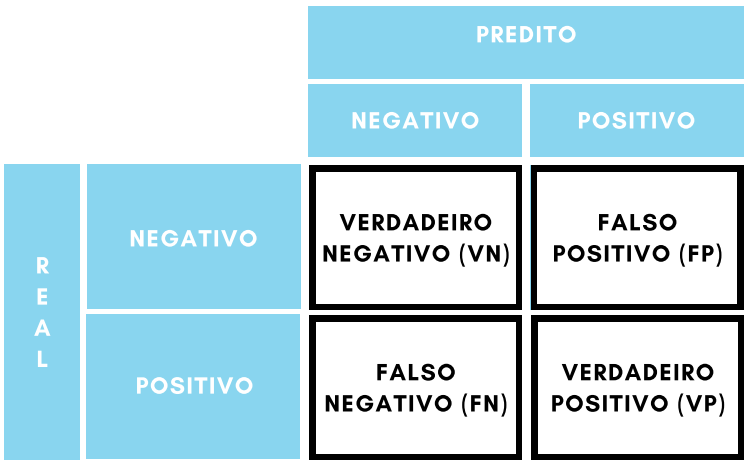

Figura 2.3: Matriz de confusão para uma classificação binária

forem representadas nas linhas, as classes reais dos dados serão representados nas colunas (ou vice e versa) (Sammut and Webb, 2011).

Um caso especial de matriz de confusão, muito utilizada em epidemiologia, ocorre quando abordamos problemas de classificação binária, onde uma das classes é dita positiva e a outra negativa, como ter ou não uma certa doença (Tchounga et al., 2014) e (Broadhurst et al., 2015). Nesse caso, cada posição da matriz ganha um nome especial (Verdadeiro negativo, falso negativo, falso positivo e verdadeiro positivo), como descrito na Figura 2.3.

A partir dessa matriz, é possível extrair diversos indicadores (Sammut and Webb, 2011), como:

- **Acurácia**: É a capacidade de identificar corretamente tanto casos positivos quanto casos negativos.

$$\frac{VP + VN}{VP + FP + VN + FN} \tag{2.1}$$

- **Especifidade**: É a capacidade de identificar corretamente, através de proporção, casos *negativos*.

$$\frac{VN}{VN + FP} \tag{2.2}$$

- **Sensibilidade**: Também conhecida como revocação ou *recall*, informa a proporção

de casos positivos identificados de maneira correta.

$$\frac{VP}{VP + FN} \tag{2.3}$$

- **Precisão**: Indica, através de proporção, a capacidade de identificar corretamente, dentre todos os casos, casos positivos.

$$\frac{VP}{VP + FP} \tag{2.4}$$

- **F1-Score**: É dada como a média harmônica entre a precisão e a sensibilidade. Varia entre 0 e 1, onde valores mais próximos a 1 indicam melhor precisão e sensibilidade.

$$F1_{\text{score}} = 2\,\frac{VP}{VP + FP + FN} = 2\,\frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.5}$$

2.5 Correlação

Como descrito em Morettin and BUSSAB (2017), a quantificação do grau de associação entre duas variáveis pode ser feita pelo chamado coeficiente de correlação, medida que descreve, por meio de um número real, a associação ou dependência entre duas variáveis. É importante destacar que a existência de correlação entre variáveis não implica, necessariamente, em causalidade (Wright, 1921).

Com o intuito de avaliar a correlação entre as variáveis da base de dados e a variável alvo, depressão, serão utilizados dois tipos de cálculos de correlação:

- Um caso especial da correlação de Pearson, chamada correlação de ponto bisserial (Lira and Neto, 2006), para cálculo de correlação entre variáveis categóricas e a variável binária, depressão.

- Correlação Cramer's V (Akoglu, 2018) para cálculo de correlação entre variáveis discretas e a variável objetivo;

É válido destacar que a correlação é calculada levando em consideração apenas os registros de valores não nulos. As tabelas 2.1 e 2.2 apresentam uma interpretação para os valores

de correlação.

Tabela 2.1: Interpretação dos valores de correlação de ponto bisserial (Akoglu, 2018).

Módulo da correlação de ponto bisserial	**Interpretação para a correlação**
0,90	muito forte
0,70	forte
0,40	moderada
0,1	fraca
0	Inexistente ou muito fraca

Tabela 2.2: Interpretação dos valores de correlação de Cramer's V (Akoglu, 2018).

Valor de Cramer's V	**Interpretação para a correlação**
0,25	muito forte
0,15	forte
0,10	moderada
0,05	fraca
0	Inexistente ou muito fraca

2.6 Psicometria

A psicometria é um campo científico da psicologia, aliado a métodos de análise estatística, que busca construir e aplicar instrumentos para mensurar variáveis de ordem psicológica (Pasquali, 2009). Existem diversos questionários psicométricos capazes de mensurar, por exemplo:

- **Depressão**: *Patient Health Questionnaire* (PHQ);
- **Dependência de nicotina**: Teste de Fagerström para a Dependência à Nicotina (FTND);
- **Transtornos de ansiedade**: Inventário de Ansiedade Traço e Estado (IDATE);
- **Dependência de álcool**: Questionário de identificação de desordem do uso de álcool (AUDIT).

O presente trabalho utiliza como ferramenta para mensurar a depressão o questionário PHQ.

Tabela 2.3: Psicometria do PHQ-9

Escala de depressão	**Intervalo de pontuação**	**Ações de tratamento propostas**
Nenhuma ou miníma	0 - 4	Nenhuma
Suave	5 - 9	Espere e repita o teste durante o acompanhamento
Moderada	10 - 15	Aconselhamento, acompanhamento e/ou farmacoterapia
Moderadamente Grave	15 - 20	Tratamento ativo com farmacoterapia e/ou psicoterapia
Grave	20 - 27	O início imediato da farmacoterapia e, em caso de comprometimento grave ou má resposta à terapia, acelerar o encaminhamento para um especialista em saúde mental para psicoterapia e/ou gestão colaborativa

[3] Tabela com os pontos de cortes e ações de tratamento relativo ao resultado do PHQ-9.

2.6.1 Questionário de Saúde do Paciente

O questionário de saúde do paciente (PHQ - *Patient Health Questionnaire*) (Kroenke et al., 2010) é uma ferramenta constituída de nove perguntas (PHQ-9), que avalia a presença dos sintomas de depressão, descritos em Batista (1995) (problemas com o sono, cansaço ou falta de energia, humor melancólico, perda de interesse ou prazer em fazer as coisas, mudança no apetite ou peso, inquietabilidade, problemas de concentração, sentimento de inutilidade ou culpa e pensamentos suicidas). As perguntas são relacionadas às duas últimas semanas e possuem quatro alternativas: nenhuma vez, vários dias, mais da metade dos dias e quase todos os dia. Cada resposta vale, respectivamente, 0, 1, 2 ou 3 pontos. Ao final das 9 perguntas é feito o somatório dos pontos relativos à cada resposta, gerando o valor do teste. A psicometria do teste segue a classificação da Tabela 2.3.

Conforme sugerido em Titov et al. (2011), nesse trabalho será utilizado pontuação de corte maior ou igual a 10 pontos como indicativo de depressão.

O PHQ também apresenta sua versão reduzida, chamada de PHQ-2, que é composta de duas perguntas. O PHQ-2 é usado como rastreio para o indicativo de depressão, tendo um desempenho favorável em relação à sua versão completa, apresentando uma especifidade de 86% e uma sensibilidade de 79% (Löwe et al., 2005) para indicar transtorno depressivo de modo geral. Richardson et al. (2010) e Staples et al. (2019) sugerem uma pontuação de corte maior ou igual a três para rastreio da depressão utilizando o PHQ-2. Caso a pontuação ultrapasse a pontuação de corte, é recomendada a aplicação

da versão completa do PHQ (Spitzer et al., 1999). O questionário PHQ, bem como os demais questionários do sistema, está disponível no apêndice A.

3 Trabalhos Relacionados

Esse capítulo visa compilar os trabalhos da literatura relacionados a esse presente estudo, a fim de entender o estado atual da área, guiando o desenvolvimento desse trabalho.

Um estudo de como a depressão impacta na cessação de consumo de tabaco pode ser encontrado em Niaura et al. (1999). O trabalho analisou diversas variáveis, como: socioeconômicas (sexo, idade, escolaridade, renda), histórico de tabagismo (número de cigarros consumidos por dia, tentativas de parar de fumar), comparando-as entre pacientes ditos depressivos e não depressivos. Um ponto importante levantado nesse trabalho é que os sintomas de depressão foram intensificados após trinta dias de abstinência.

Covey et al. (1990) discorre sobre pacientes depressivos terem menor taxa de sucesso na cessação do consumo de tabaco que pacientes não depressivos, mostrando que pacientes depressivos têm intensificação dos sintomas da depressão e apresentam dificuldades de concentração durante a privação de nicotina.

Yang et al. (2016) desenvolveu modelos de árvore de decisão para classificação de depressão segundo os dados de distribuição do PHQ-8 (uma adaptação do PHQ-9 (Kroenke et al., 2009)). O modelo foi dividido em dois, onde foi gerado uma árvore de decisão para o sexo masculino e outra para o sexo feminino. Apesar de mostrar resultados promissores, não é possível generalizar os resultados obtidos devido a sua pequena amostra de teste.

Um modelo de predição de diagnóstico de dengue utilizando árvores de decisão foi desenvolvido por Tanner et al. (2008). Os resultados obtiveram 71.2% de sensibilidade e 90.1% de especifidade sobre uma amostra de 1200 pacientes de Singapura e Vietnã. A principal vantagem do modelo de árvore de decisão foi a inteligibilidade das tomadas de decisões para o diagnóstico do paciente, levando ao entendimento dos fatores que diferenciam uma febre causada por outros fatores daquela causada pela dengue.

Al Jarullah (2011) gerou modelos de árvores de decisão, utilizando o algoritmo J48, para diagnóstico de diabetes do tipo II a partir de dados da base PIMA (Pima Indians Diabetes Database) (Rossi and Ahmed, 2015). O modelo levou em conta variáveis como:

pressão sanguínea, índice de massa corporal e idade dos pacientes, sendo capaz de replicar decisões já conhecidas da medicina, como a associação entre obesidade e diabetes tipo II.

Patel et al. (2016) faz uma revisão da literatura mostrando diversos trabalhos que utilizaram modelos de máquina de vetor suporte para detectar a presença de depressão em pacientes, através da análise de imagens de exames de ressonância magnética. Os resultados apresentados mostram-se promissores.

Kessler et al. (2016) utiliza um conjunto de árvores de decisão para detectar depressão através de dados de um estudo transversal retrospectivo, onde o próprio paciente reporta as respostas. O resultado obtido se mostra superior quando comparado a técnicas já utilizadas na área, como a regressão logística.

4 Base de Dados do Viva Sem Tabaco

Esse capítulo descreve na secção 4.1 quais são os dados utilizados nesse trabalho, bem como eles foram obtidos, quais foram os critérios utilizados para composição da base. A secção 4.2 apresenta a caracterização dos usuários presentes na base de dados. A secção 4.3 mostra uma análise das variáveis e, por fim, a secção 4.4 expõe uma análise de correção das variáveis com a depressão.

4.1 Obtenção dos dados

Os dados utilizados foram extraídos do banco de dados do Viva Sem Tabaco. O banco contém informações socioeconômicas (como idade, sexo, nível de formação acadêmica) e respostas de questionários realizados na intervenção, como o questionário sobre a saúde do paciente, (Kroenke et al., 2010) (PHQ, do inglês Patient Health Questionnaire), usado para detectar o grau de depressão do paciente, e o Teste de Fagerström para dependência de nicotina. Heatherton et al. (1991) (FTND, Fagerström Test for Nicotine Dependence).

O sistema do Viva Sem Tabaco é dividido em três etapas: pronto para parar (também chamada de vale a pena parar?), preparando para parar (ou quero parar) e acompanhamento (ou já parei), como pode ser visto na Figura 4.1. A primeira etapa, chamada de preparando para parar, acontece com o usuário ainda não logado. Nessa etapa é apresentado material informativo sobre o consumo de tabaco, e são aplicados questionários de rastreio de depressão (PHQ-2), dependência de tabaco (FTND) e de dependência de álcool (AUDIT) Bush et al. (1998).

Ao final dessa etapa, o usuário é convidado a criar uma conta no sistema. Durante a segunda etapa, pronto para parar, são apresentados outros materiais informativos e ocorre a construção do plano para parar, onde o usuário define uma data e um plano de ação para cessar o consumo de tabaco. Nessa etapa o usuário volta a responder o questionário de dependência de tabaco e a uma versão estendida do PHQ, o PHQ-9.

A terceira etapa, acompanhamento, é utilizada por aqueles usuários que já pos-

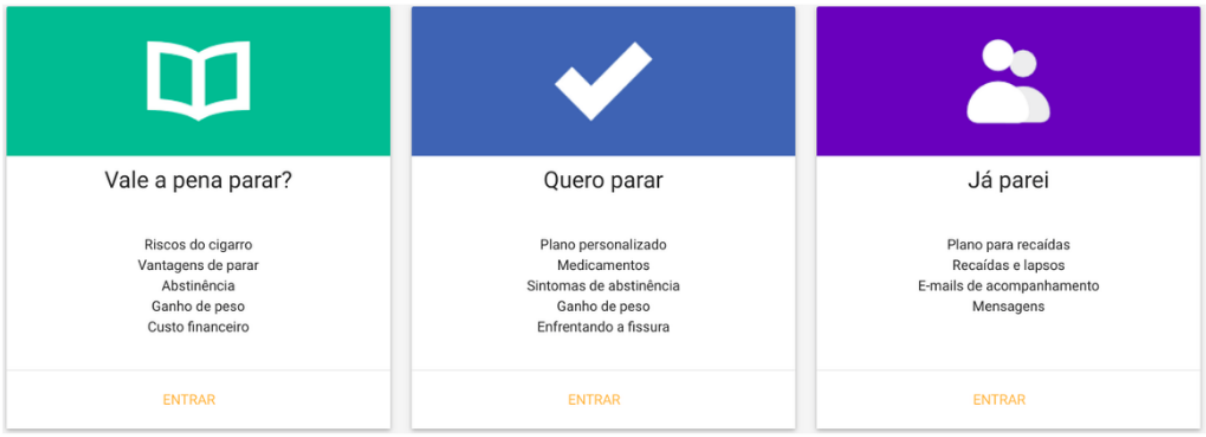

Figura 4.1: Etapas da intervenção online do Viva Sem Tabaco.

suem uma conta no sistema e já pararam de fumar, ou que já criaram um plano para cessação do consumo da droga. Essa etapa visa conseguir o feedback dos usuários que conseguiram parar de fumar utilizando o plano e possibilitar a criação de um novo plano para aqueles que tiveram uma recaída.

O banco contém 1756 usuários, que se cadastraram no sistema entre 22/01/2014 e 01/03/2018. Dos usuários, 945 responderam algum dos questionários da primeira etapa sistema, pronto para parar. Quando é avaliado o número de usuários que responderam a todos questionários dessa etapa, o número cai para 315. Sobre o total de usuários foi aplicado uma lista de exclusão, criada manualmente, com a finalidade de remover usuários de teste (criados pela equipe de desenvolvimento do sistema), usuários falsos ou, de modo geral, usuários que não se qualificavam para a pesquisa, como os que não aceitaram participar da pesquisa ao se cadastrarem no sistema. Após essa filtragem, foram selecionados 428 usuários para compor a base de dados utilizada nesse trabalho.

4.2 Caracterização da Base de Dados

A fim de caracterizar qual é o perfil do usuário que utiliza a intervenção online do Viva Sem Tabaco, a Tabela 4.1 apresenta uma caracterização dos usuários. 68% dos usuários são mulheres. A idade dos usuários variam de 18 a 70 anos, com média de 43 anos e desvio

Tabela 4.1: Características sociodemográficas dos usuários cadastrados no Viva Sem Tabaco

Variáveis		%	**N**
Sexo			428
Homens	137	32,00%	
Mulheres	291	68,00%	
Idade			428
Máximo	70		
Mínimo	18		
Média	43.19		
Mediana	43		
Desvio Padrão	11.58		
Escolaridade			186
Ensino Fundamental	16	8,60%	
Ensino Médio	76	40,86%	
Ensino Superior	69	37,10%	
Pós Graduação	25	13,44%	
Trabalhando			186
Sim	124	66,33%	
Não	62	33,67%	
Depressivos			428
Sim	178	41,59%	
Não	250	58,41%	

padrão de 11,59. A maior parte dos usuários, 40,86%, possui ensino médio, seguido de 37,10% que possuem ensino superior. Um terço dos usuários, 66,67%, está trabalhando atualmente.

4.3 Análise dos dados

A base de dados contém 428 registros e um total de 44 variáveis. Na Tabela B.1 é possível conferir uma lista das variáveis e dos possíveis valores assumidos por elas.

Serão utilizadas as respostas do questionário psicométrico PHQ-9, descrito na sessão 2.6.1. As variáveis phq_1, phq_2, phq_3, phq_4, phq_5, phq_6, phq_7, phq_8 e phq_9, respostas de cada uma das questões do teste, são somadas para gerar uma variável chamada phq_score. Com base na variável phq_score, é utilizada a pontuação de corte, maior ou igual a dez, descrita na Tabela 2.3, para indicativo de depressão.

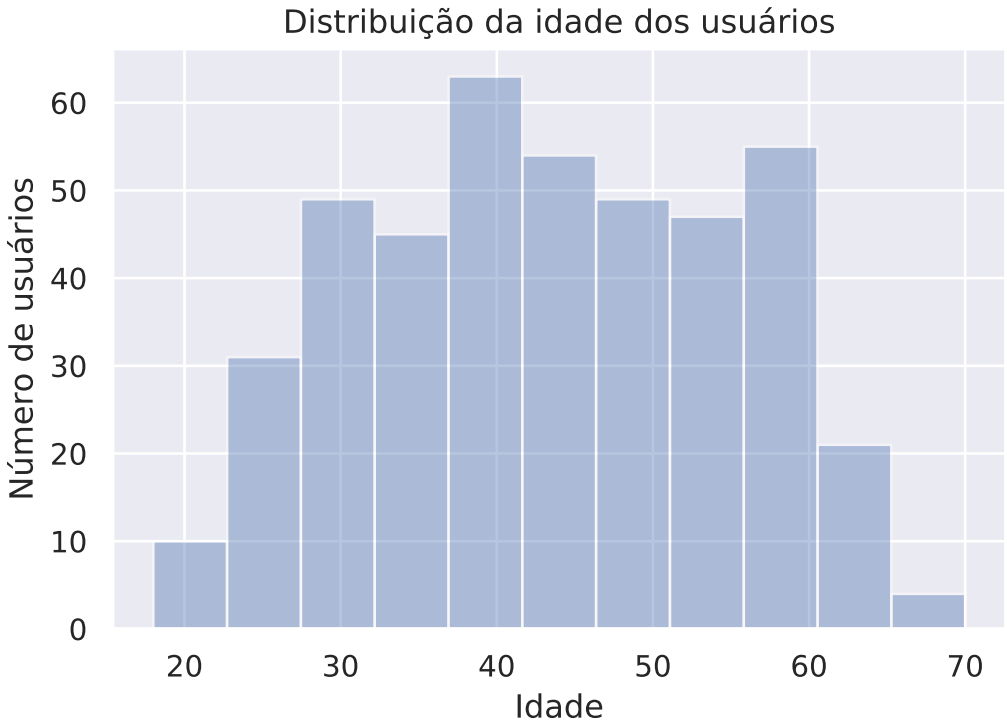

Figura 4.2: Distribuição da idade dos usuários.

4.3.1 Análise preliminar das variáveis

Analisando a Figura 4.2, é possível perceber que existe uma grande concentração de usuários com mais de 40 anos, o que concorda com os dados apresentados em Cohen and Lichtenstein (1990).

Com base na Figura 4.3, é possível perceber que o maior número de usuários apresentam ensino médio ou superior completos. Também é possível notar que as mulheres usuárias do Viva Sem Tabaco, têm, em média, maior grau de escolaridade se comparadas aos homens, com valores percentuais de pós-graduações mais que três vezes maior que os presentes em usuários do sexo masculino.

A Figura 4.4 mostra que a maior partes dos usuários está empregada. Também é possível notar uma leve diferença entre a porcentagem de homens e mulheres empregados, onde a porcentagem de homens empregados é em torno de 10% maior que a porcentagem de mulheres trabalhando.

Analisando a Figura 4.5, é possível perceber que a maior parte dos usuários (cerca de 80%) informaram que já tentaram parar de fumar ao menos uma vez.

Conforme ilustrado na Figura 4.6, a maioria dos usuários indicaram desejo parar

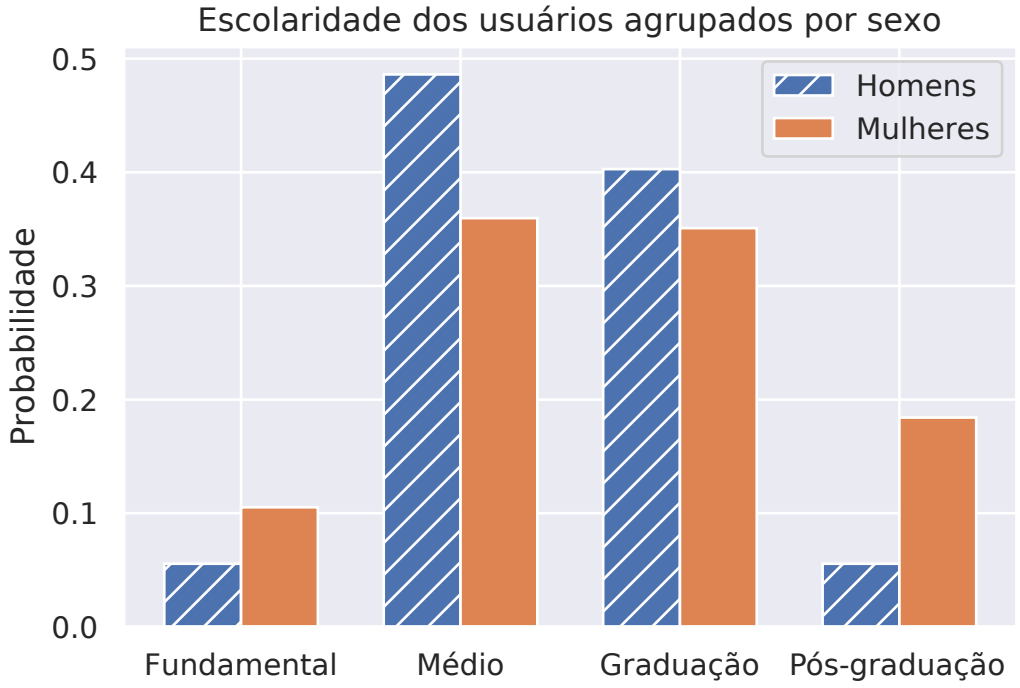

Figura 4.3: Distribuição da escolaridade dos usuários, separados por sexo.

de fumar em até 7 dias, a partir da criação do seu plano de cessação de consumo de tabaco.

A Figura 4.7 e a Tabela 4.2 mostram a distribuição da variável ladder, indicador de motivação da cessação do consumo de tabaco. Os valores variam de 0 (nenhuma motivação) até 10 (extremamente motivado). Podemos notar uma maior concentração em torno de 8, indicando uma alta motivação para cessação do consumo de tabaco.

A Figura 4.8 foi criada utilizando as repostas do teste PHQ-9 e usando a pon-

Tabela 4.2: Mostrando os possíveis valores acompanhados do seu respectivo significado para a variável ladder. Quanto maior o valor, mais motivado o usuário está para parar de fumar.

Valor	**Opção**
10	Eu já desisti e não vou mais fumar
9	Eu parei, mas ainda estou preocupado com a recaída. Eu ainda preciso trabalhar para ficar parado
8	Eu ainda fumo, mas comecei a reduzir o número de cigarros que fumo
7	Tenho planos para sair dentro de 30 dias
6	Tenho planos para sair dentro de 6 meses
5	Muitas vezes penso em parar de fumar, mas não tenho planos
4	Às vezes penso em desistir, mas não tenho planos de desistir
3	Eu raramente penso em desistir e não tenho planos para desistir
2	Eu nunca penso em desistir e não tenho planos de desistir
1	Eu gosto de fumar e não considero parar

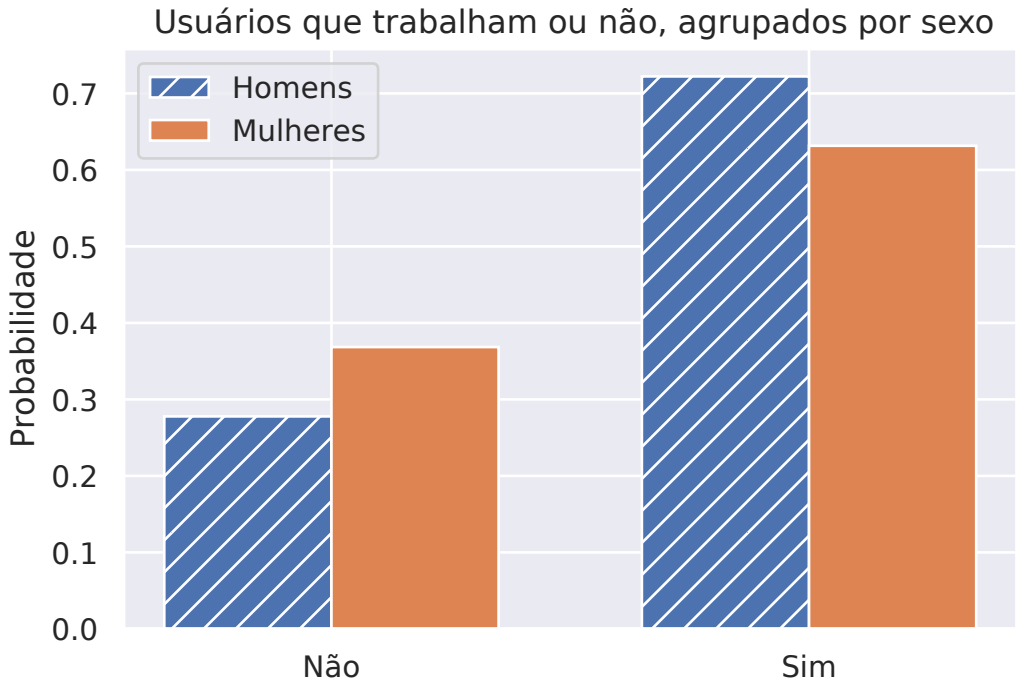

Figura 4.4: Distribuição de usuários trabalhando, separados por sexo.

tuação de corte igual a 10 para indicativo de depressão, conforme sugerido por Titov et al. (2011). É possível notar que a maior partes dos usuários, quase 60%, foi classificada como não depressiva, segundo ao PHQ-9. Não foi possível perceber uma diferença expressiva na presença de depressão entre homens e mulheres.

4.4 Análise de correlação

Analisando a Figura 4.9, com base na interpretação da correlação segundo a Tabela 2.2, existe uma correlação moderada das variáveis employed e enfrentar_fissura_ler_razoes com a depressão, porém não foi encontrada nenhuma outra correlação mais expressiva. Já utilizando a Tabela 2.1, para analisar a Figura 4.10, não foi encontrada nenhuma correlação linear, moderada ou forte, entre as variáveis discretas e a variável objetivo, sugerindo que os dados possam contar com correlações fortemente não lineares, ou possivelmente não apresentarem correlação com a depressão.

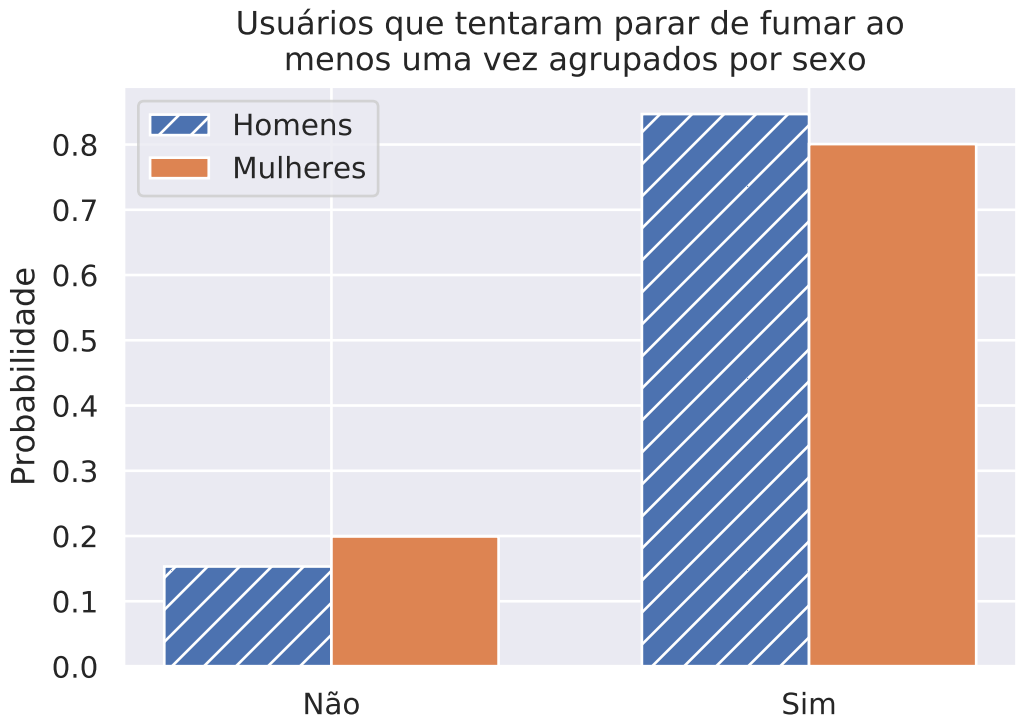

Figura 4.5: Variável tentou parar dividida por sexo.

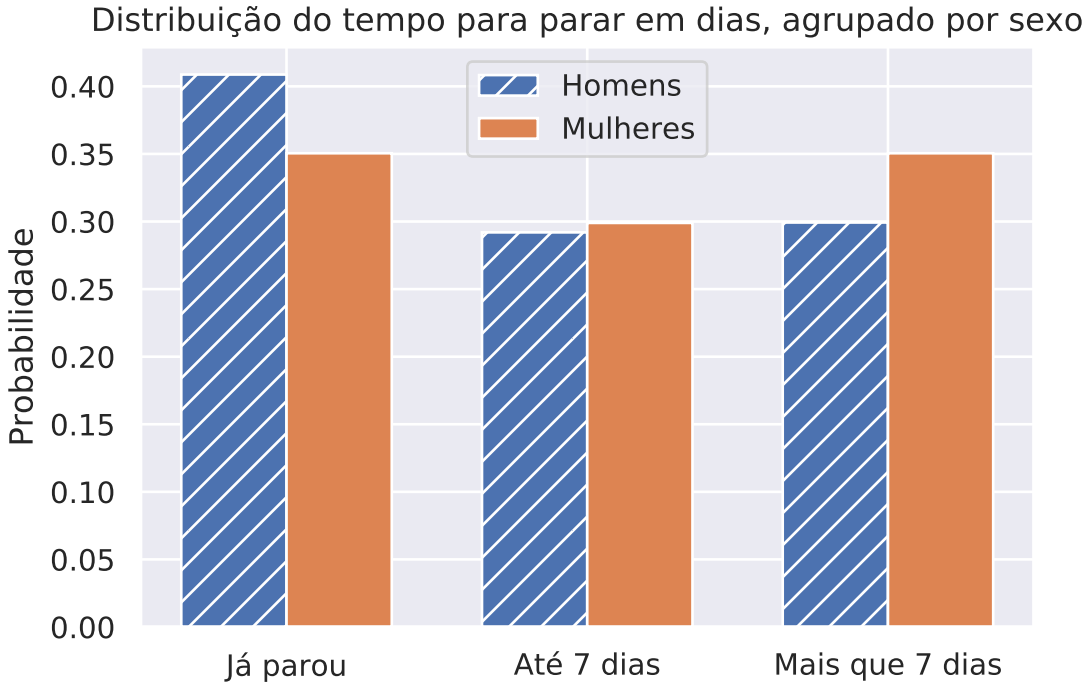

Figura 4.6: Distribuição variável tempo para parar.

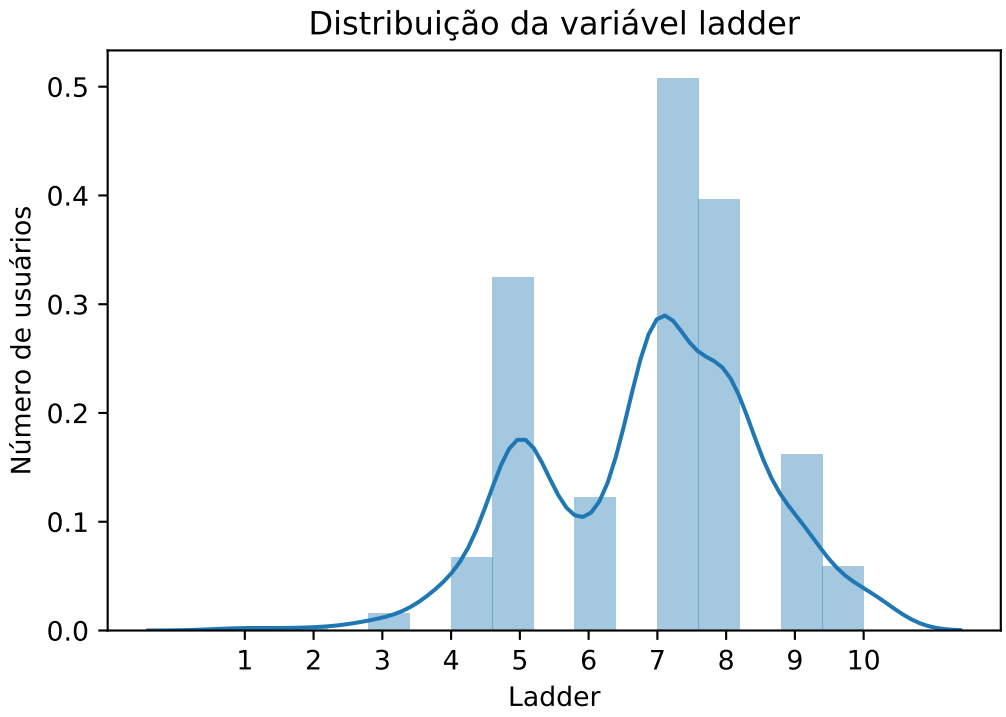

Figura 4.7: Distribuição da variável ladder.

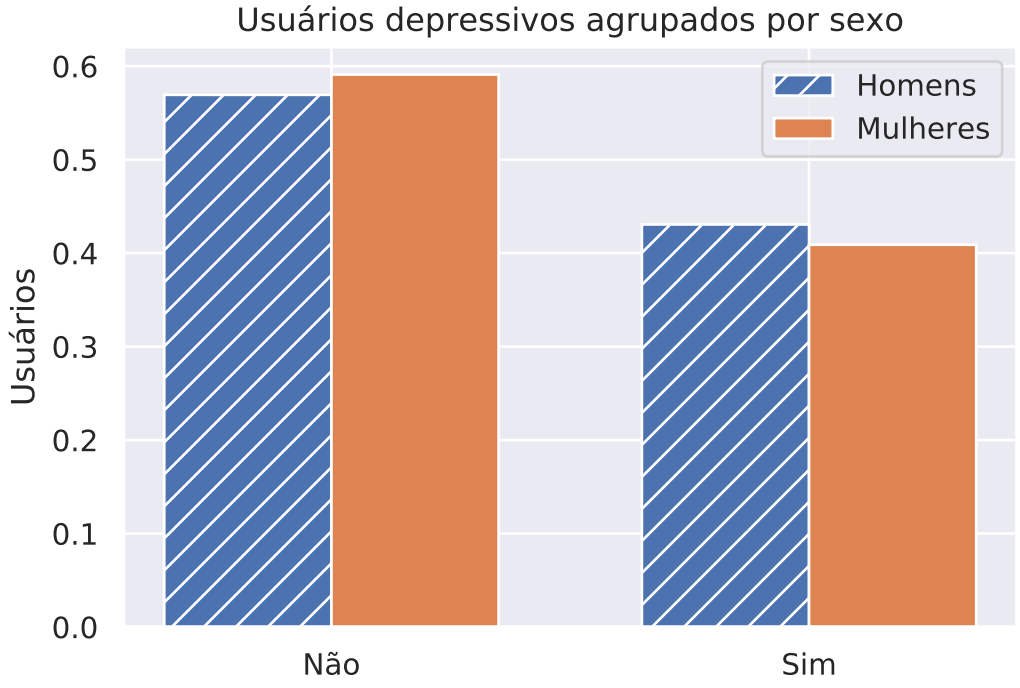

Figura 4.8: Distribuição dos usuários depressivos.

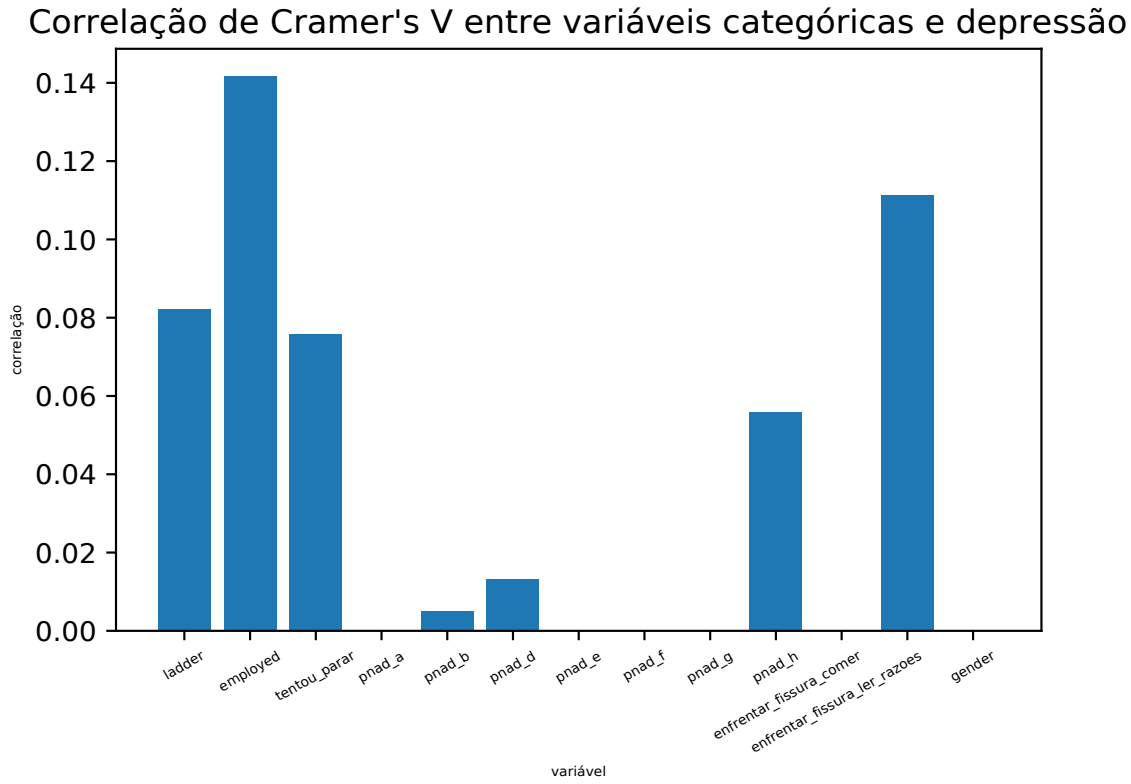

Figura 4.9: Correlação entre as variáveis categóricas e a variável depressão.

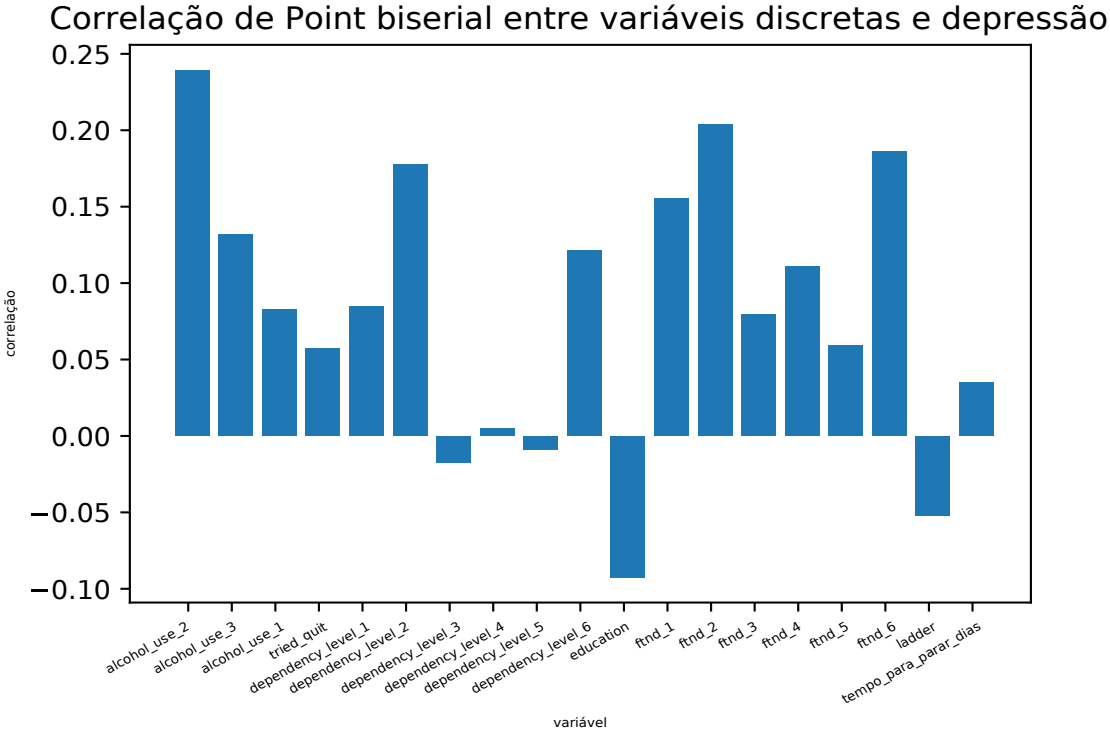

Figura 4.10: Correlação entre as variáveis discretas e a variável depressão.

5 Modelos para Predição de Depressão

Esse capítulo apresenta a descrição do processo de geração dos modelos, bem como os parâmetros utilizados e, finalmente, uma breve discussão da interpretação dos modelos e de suas métricas.

O procedimento adotado nesse trabalho consiste em:

- Preparação dos dados;
- Treinamento do modelo de árvore de decisão de repartição binária (RPART);
- Avaliação do modelo segundo as métricas descritas na seção 2.4;
- Estudo do modelo e análise de relação das variáveis.

Para treinamento do modelo foi utilizada a estratégia de validação cruzada estratificada, k-folds (k=10) (Zeng and Martinez, 2000). A altura máxima da árvore de decisão foi limitada em 8 níveis, a fim de garantir um modelo de melhor interpretabilidade. Também foi utilizado um peso para balanceamento das classes. O peso para a classe i é calculado conforme a equação 5.1, em que N_{TOTAL} é o número total de amostras na base de dados e Y_i é o número de amostras da classe i.

$$Peso_i = \frac{N_{\text{TOTAL}}}{Y_i} \tag{5.1}$$

Para a geração dos modelos obtidos no presente trabalho, a base de dados, descrista na secção 4 (N=428), foi divididas de 3 maneiras:

- Removendo todos os registros com valores faltantes (N=186)
- Removendo todas as colunas de valores faltantes (N=428)
- Utilizando todos os registros (N=428)

5.1 Modelos Gerados Removendo Valores Faltantes

Para gerar o primeiro modelo, foi utilizado apenas os registros sem valores faltantes, totalizando 186 amostras, divididas entre 74 depressivos (39,78%) e 112 não depressivos (60,22%), segundo o PHQ-9. O modelo obtido pode ser visto na Figura 5.1 e suas métricas para detecção de depressão são encontradas na Tabela 5.1. Cada nó da árvore é composto de três valores. O valor da primeira linha representa a classe com maior peso naquele nó (1 = depressivo, 0 = não depressivo), o valor do meio representa a probabilidade dos dados daquele nó pertencerem a classes 1 (depressivo) e o valor da última linha representa a porcentagem de registros presentes naquele nó.

Analisando o modelo, podemos perceber que o principal fator que influenciou a tomada de decisão para diagnóstico de depressão foi a quantidade de álcool consumida, onde pacientes que relataram ingerir álcool duas ou mais vezes mensalmente apresentaram maiores taxas de depressão. Também foi possível relacionar a quantidade de tentativas de parar de fumar e desemprego com a depressão, observando que:

- Usuários que realizam consumo pesado de álcool e tentaram parar mais de 4 vezes são depressivos em 83% dos casos;
- Usuários que realizam consumo pesado de álcool, tentaram parar menos de 5 vezes e estão desempregados são considerados depressivos com 82% de certeza.

Outro ponto de destaque foi que usuários com baixo consumo de álcool que têm ao menos ensino médio e que relataram não fumar nos 5 primeiros minutos após acordarem, são ditos **não** depressivos com 77% de certeza.

Tabela 5.1: Métricas do modelo apresentado na Figura 5.1.

N	**Acurácia**	**Precisão**	**Sensibilidade**	**Especifidade**	**F1-Score**
186	75,81%	68%	74%	78%	71%

Analisando os resultados que indicaram relação do consumo de álcool e depressão, foi proposto uma modificação na base de dados, a fim de investigar a relação entre o consumo de álcool e depressão. Foi realizada a transformação de três variáveis presentes na base de dados, que representam as respostas do questionário de dependência de álcool,

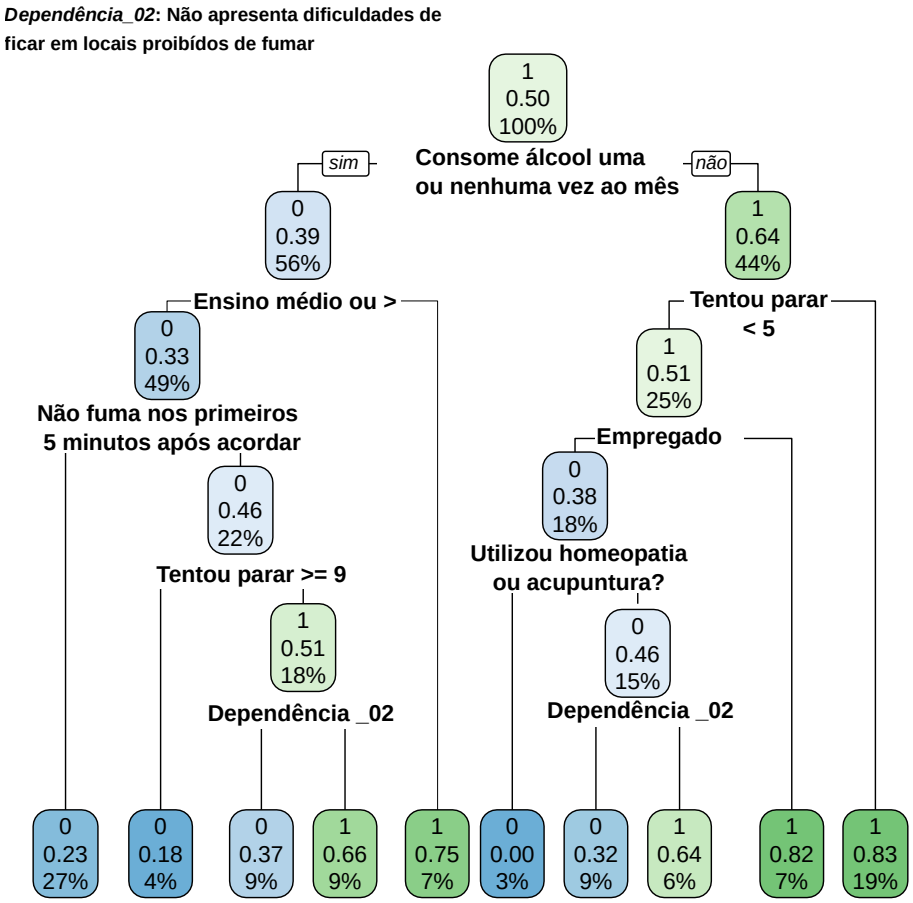

Figura 5.1: Modelo de predição de depressão utilizando apenas registros sem valores faltantes.

AUDIT, em uma nova variável que indica se o nível do consumo de álcool traz malefícios à saúde (Bush et al., 1998). Foi feita a soma das pontuações de cada uma das três perguntas do AUDIT e, posteriormente, foi aplicada uma pontuação de corte maior que 6 para mulheres e maior que 7 para homens, como indicativo do consumo prejudicial. O modelo obtido é apresentado com sua matriz de confusão nas figuras: Figura 5.3 e Figura 5.4 respectivamente. As métricas do modelo para detecção de depressão são encontradas na Tabela 5.2. Realizando a análise do modelo obtido, foi possível perceber que a nova variável só apareceu em níveis mais inferiores da árvore, indicando menor relevância. A variável que apresentou maior importância para o novo modelo foi a idade. As principais tomadas de decisão a serem destacas nesse modelo são:

- Usuários que têm idade menor que 47 anos, estão desempregados e fumam o primeiro cigarro logo nos 30 primeiros minutos após acordar são depressivos, com 85% de certeza.

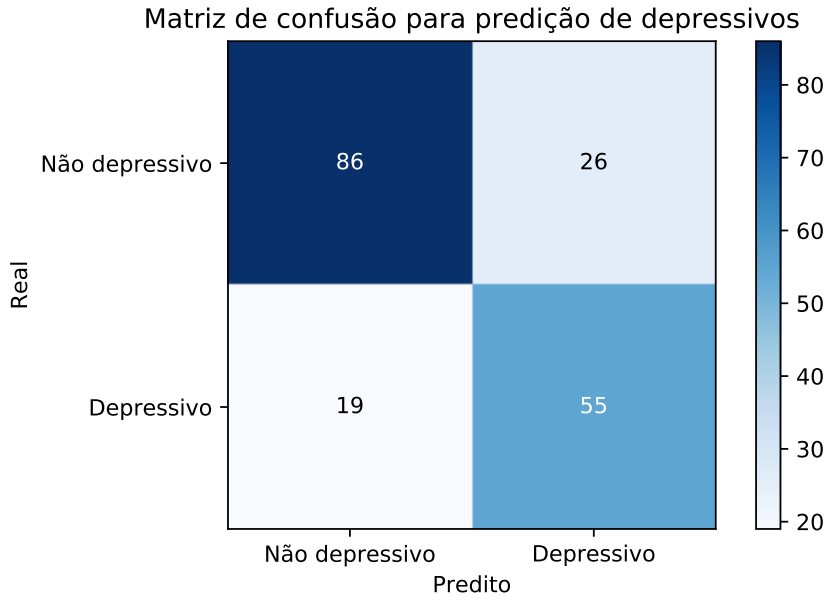

Figura 5.2: Matriz de confusão do modelo mostrado na Figura 5.1.

Tabela 5.2: Métricas do modelo apresentado na Figura 5.3.

N	Acurácia	Precisão	Sensibilidade	Especifidade	F1-Score
186	78,50%	69%	82%	76%	75%

- Usuários entre 29 e 47 anos, que encontram-se empregados e relataram ter dificuldades de permanecerem em locais que são proibidos de fumar (como escolas, igrejas e hospitais) são considerados depressivos, com 73% de certeza.

Comparando as métricas do modelo da Figura 5.1 e do modelo da Figura 5.3, é possível notar uma diminuição dos valores de falso negativo (de 19 registros para 13 registros), além de uma melhora na acurácia obtida (de 75,81% para 78,50%). Logo, a modificação das variáveis do questionário AUDIT tiveram impacto positivo nas métricas do modelo.

5.2 Modelos Gerados Removendo variáveis com valores faltantes

Em uma segunda análise foram removidas todas as colunas com atributos faltantes obtendo 428 registros e 20 variáveis. Dos usuários, 178 (41.59%) eram depressivos e 250 (58.41%) eram não depressivos, segundo o PHQ-9. O modelo obtido e sua matriz de

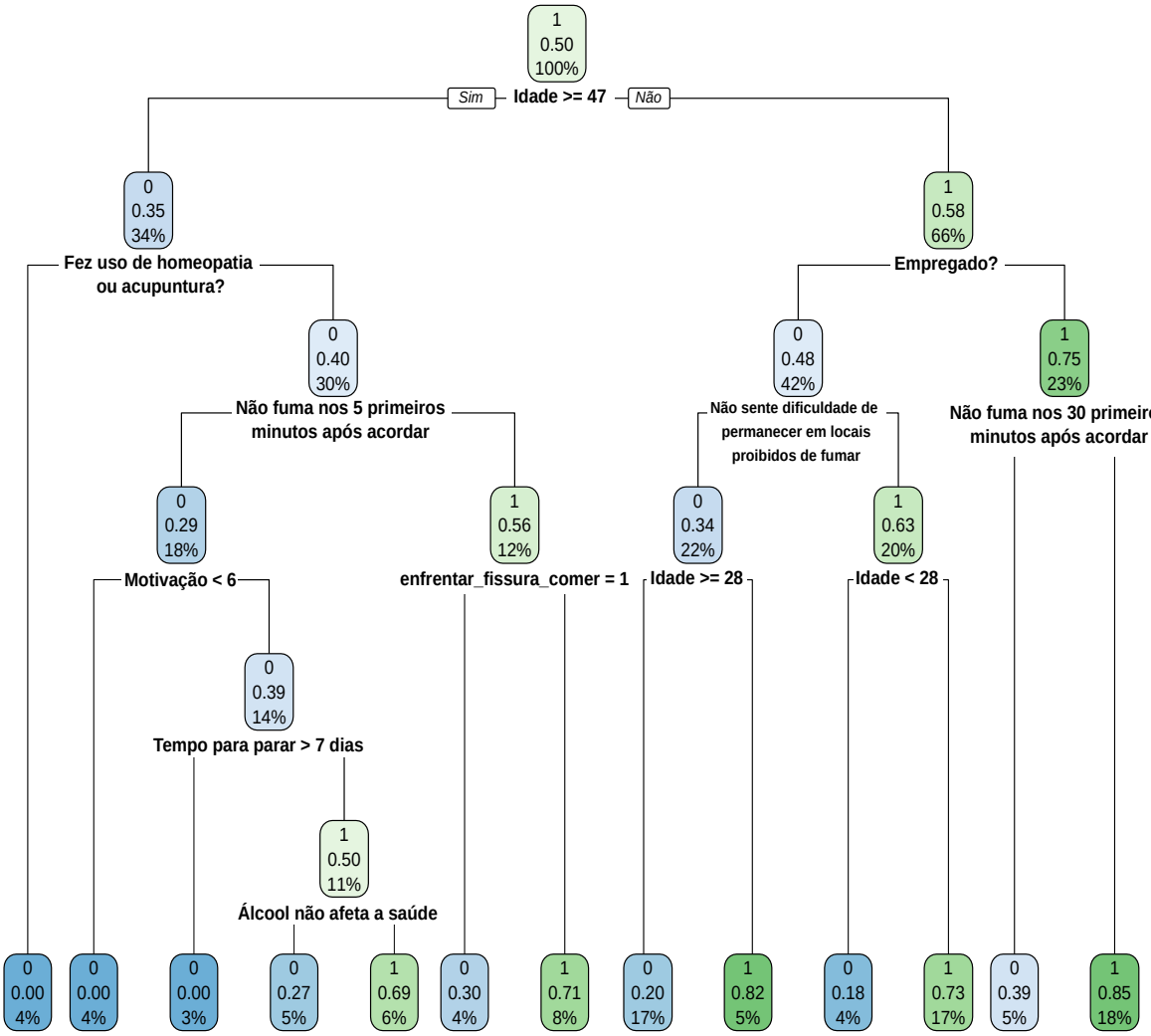

Figura 5.3: Modelo de predição de depressão utilizando apenas registros sem valores faltantes e aplicando a transformação nas variáveis do questionário AUDIT.

confusão podem ser visto nas figuras 5.5 e 5.6. As métricas do modelo são encontradas na Tabela 5.2. As principais tomadas de decisão para esse modelo foram:

- Se o usuário **não** tem dificuldades de permanecer em locais em que é proibido fumar, tem idade idade superior a 27 anos, **não** considera o primeiro cigarro da manhã o mais difícil de desistir e consome menos de 24 cigarros ao dia, então ele é dito não depressivo, com 77% de certeza;

- Se o usuário tem dificuldade de permanecer em locais em que é proibido fumar e tem idade entre 29 e 58 anos, então ele é dito depressivo, com 70% de confiança.

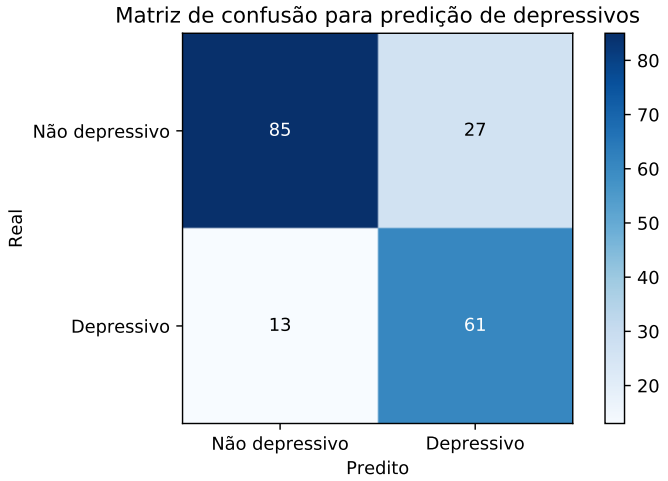

Figura 5.4: Matriz de confusão do modelo mostrado na figura 5.3

Tabela 5.3: Métricas do modelo apresentado na Figura 5.5.

N	Acurácia	Precisão	Sensibilidade	Especifidade	F1-Score
428	70,56%	62%	75%	68%	68%

5.3 Modelos Gerados Utilizando Todos os Registros e Variáveis

Utilizando todos os 428 registros e 33 variáveis temos que, 178 (41.59%) dos usuários eram depressivos e 250 (58.41%) eram não depressivos, segundo o PHQ-9. O modelo obtido e sua matriz de confusão podem ser vistos nas Figuras 5.7 e 5.8. As métricas do modelo são encontradas na Tabela 5.4. O modelo encontrado foi mais complexo que os anteriores. Suas principais tomadas de decisão são:

- Usuários que **não** têm dificuldades de fumar em locais proibidos, têm idade maior que 27 anos, estão empregados e relataram comer como método de combate a fissura, são ditos **não** depressivos, com 80% de certeza.

- Usuários que têm dificuldades de fumar em locais proibidos, possuem idade entre 28 e 58 anos, fazem consumo de álcool até três vezes por semana e declararam querer parar de fumar em até 7 dias depois da criação do plano de cessação do consumo

Tabela 5.4: Métricas do modelo apresentado na Figura 5.7.

N	Acurácia	Precisão	Sensibilidade	Especifidade	F1-Score
428	73,60%	67%	72%	69%	70%

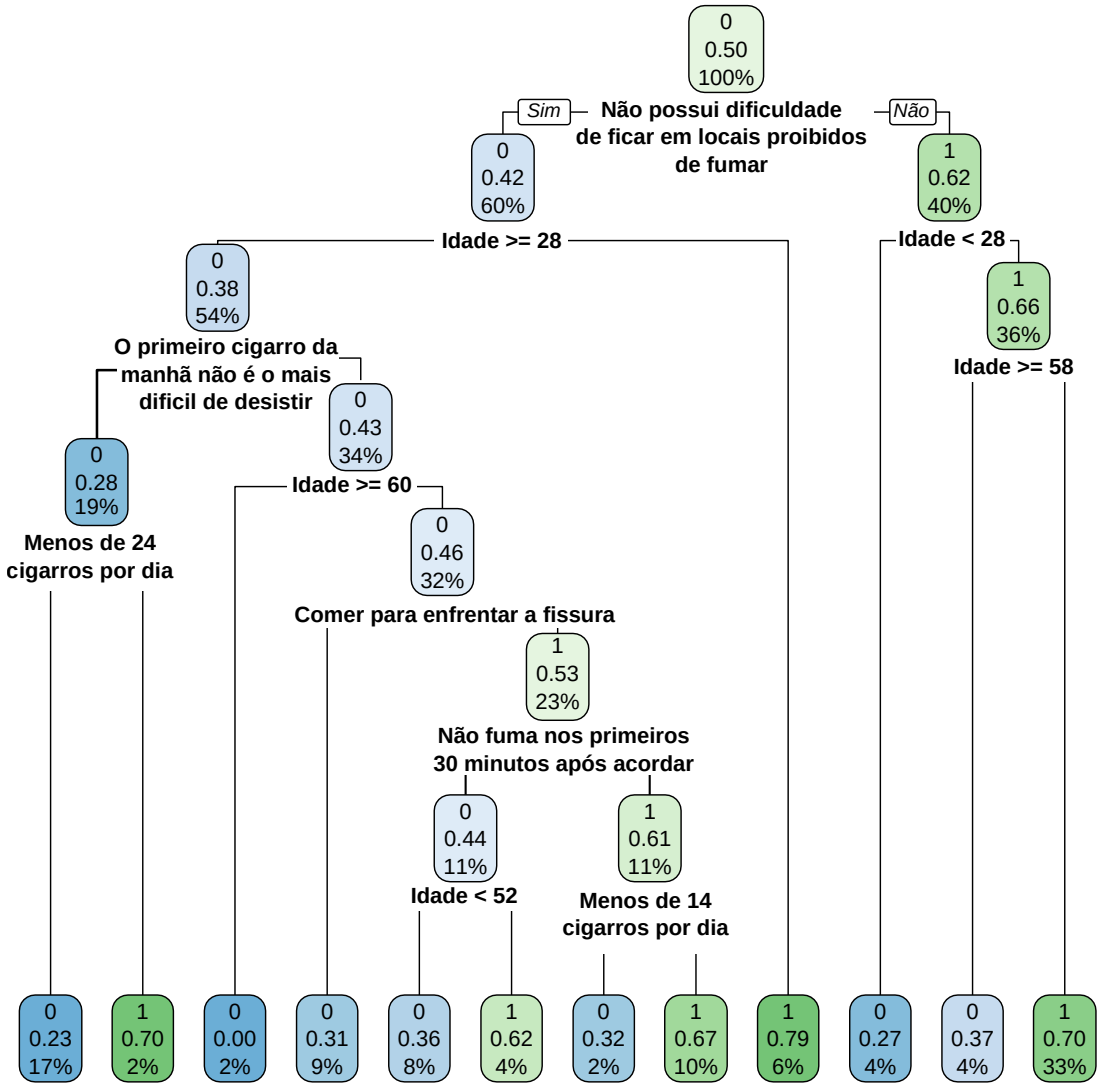

Figura 5.5: Modelo de predição de depressão utilizando todos registros e removendo colunas com valores faltantes.

de tabaco, são ditos depressivos, com 81% de certeza.

- Usuários que têm dificuldades de fumar em locais proibidos, possuem idade entre 28 e 45 anos, fazem consumo de álcool até três vezes por semana e declararam querer parar de fumar em mais de 7 dias depois da criação do plano de cessação do consumo de tabaco, são ditos depressivos, com 68% de certeza.

Comparando os modelos apresentados nas figuras: Figura 5.5 e Figura 5.7, foi possível notar que o maior número de variáveis presentes no modelo da Figura 5.7 proporcionou um melhor desempenho, apresentando melhora em quatro das cinco métricas descritas.

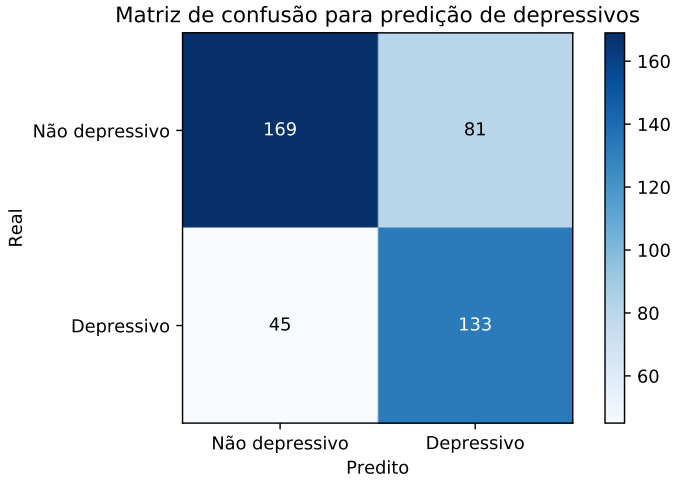

Figura 5.6: Matriz de confusão do modelo mostrado na figura 5.5.

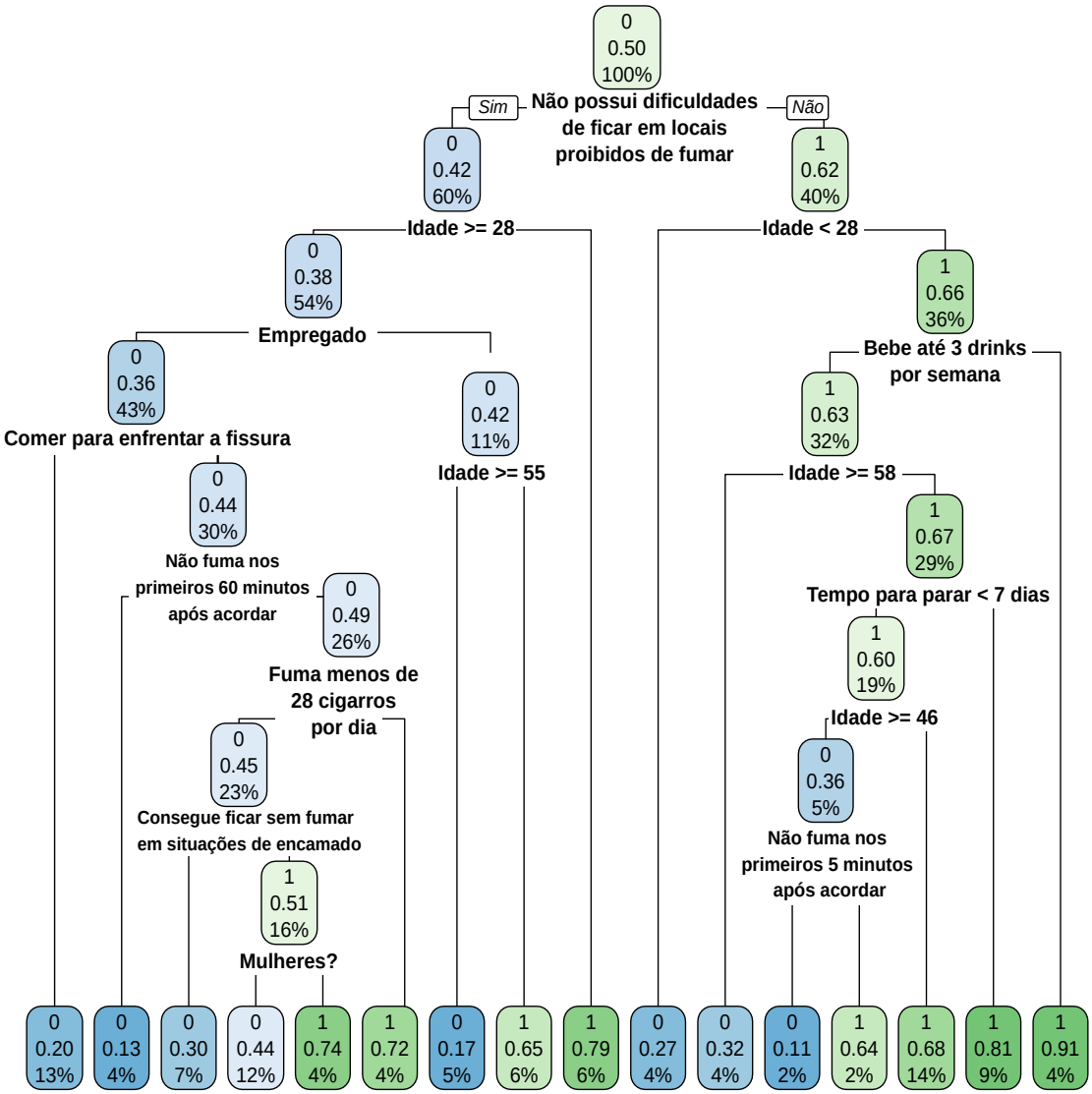

Figura 5.7: Modelo de predição de depressão utilizando todos registros.

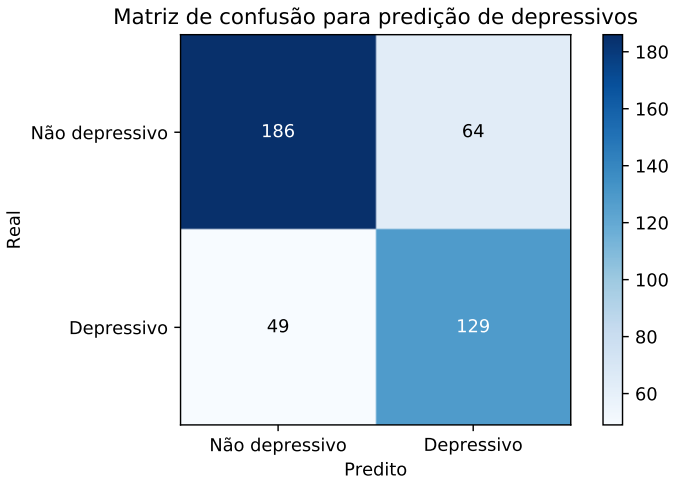

Figura 5.8: Matriz de confusão do modelo mostrado na figura 5.7.

6 Conclusões e Trabalhos Futuros

O presente trabalho apresentou uma análise da base de dados do Viva Sem Tabaco, bem como a distribuição dos seus principais atributos. Também foi realizada uma caracterização dos usuários da intervenção, a fim de entender quem são seus usuários. Posteriormente foram gerados modelos de predição de depressão, utilizando árvores de decisão.

Com base na caracterização dos usuários apresentada, foi possível notar que a maior parte dos usuários da intervenção (68%) são mulheres. A média da idade foi de 43,19 anos. Dos usuários, 40,68% possuem ensino médio e 37,10% possuem ensino superior. Do total 66,31% encontram-se empregados. Segundo o teste (PHQ-9), aplicado durante a utilização do sistema, 41,59% dos usuários foram avaliados como depressivos.

Através dos modelos gerados, foi possível observar uma relação entre as variáveis, consumo de álcool, dependência de nicotina, idade e a depressão. Outro ponto observado foi a maior taxa de depressão em usuários desempregados. Porém, não é possível indicar se há causalidade e, em caso positivo, quem é a causa e quem é o efeito nessa última observação. Em outras palavras, não é possível dizer se os usuários estão depressivos por estarem desempregados, se estão desempregados por estarem depressivos, ou se não há uma associação direta. Com base nas análises realizadas, foi possível perceber que utilizar o maior número de variáveis para a geração da árvore de decisão impactou positivamente o desempenho do modelo. Também é valido destacar que a transformação das variáveis do questionário AUDIT se mostrou útil na diminuição de falsos negativos, além de proporcionar uma melhora na acurácia do modelo.

Como trabalhos futuros, pretende-se utilizar outras técnicas de aprendizado de máquina para a geração de modelos, como a programação genética. Também deseja-se incorporar os modelos obtidos no sistema do Viva Sem Tabaco, a fim de melhorar a qualidade da intervenção, oferecendo tratamento personalizado a usuários ditos depressivos, a fim de maximizar a taxa de cessação desse grupo.

Bibliografia

Aggarwal, C. C. (2014). *Data classification: algorithms and applications*. CRC press.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*.

Al Jarullah, A. A. (2011). Decision tree discovery for the diagnosis of type ii diabetes. In *2011 International conference on innovations in information technology*, pages 303–307. IEEE.

Augusto, D. A. (2009). *Programaçao Genética Multi-populacional e Co-evolucionária para Classificaçao de Dados*. PhD thesis, Universidade Federal do Rio de Janeiro.

Batista, D. (1995). *Manual diagnóstico e estatístico de transtornos mentais*. Artes Médicas.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. wadsworth int. *Group*, 37(15):237–251.

Broadhurst, M. J., Kelly, J. D., Miller, A., Semper, A., Bailey, D., Groppelli, E., Simpson, A., Brooks, T., Hula, S., Nyoni, W., et al. (2015). Reebov antigen rapid test kit for point-of-care and laboratory-based testing for ebola virus disease: a field validation study. *The Lancet*, 386(9996):867–874.

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., and Bradley, K. A. (1998). The audit alcohol consumption questions (audit-c): an effective brief screening test for problem drinking. *Archives of internal medicine*, 158(16):1789–1795.

Cohen, S. and Lichtenstein, E. (1990). Partner behaviors that support quitting smoking. *Journal of consulting and clinical psychology*, 58(3):304.

Covey, L. S., Glassman, A. H., and Stetner, F. (1990). Depression and depressive symptoms in smoking cessation. *Comprehensive Psychiatry*, 31(4):350–354.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Garcia, S. C. (2003). O uso de árvores de decisão na descoberta de conhecimento na área da saúde.

Gomide, H., Bernardino, H. S., Richter, K., Martins, L., and Ronzani, T. (2016). Development of an open-source web-based intervention for brazilian smokers–viva sem tabaco. *BMC medical informatics and decision making*, 16(1):103.

Gomide, H. P. et al. (2017). Viva sem tabaco-características dos usuários de uma intervenção mediada por internet para fumantes.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., and FAGERSTROM, K.-O. (1991). The fagerström test for nicotine dependence: a revision of the fagerstrom tolerance questionnaire. *British journal of addiction*, 86(9):1119–1127.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., Ebert, D. D., Hwang, I., Li, J., de Jonge, P., et al. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular psychiatry*, 21(10):1366.

Khan, M. E., Khan, F., et al. (2012). A comparative study of white box, black box and grey box testing techniques. *Int. J. Adv. Comput. Sci. Appl*, 3(6).

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Kroenke, K. and Spitzer, R. L. (2002). The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515.

Kroenke, K., Spitzer, R. L., Williams, J. B., and Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General hospital psychiatry*, 32(4):345–359.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. (2009). The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Lira, S. A. and Neto, A. C. (2006). Coeficientes de correlação para variáveis ordinais e dicotômicas derivados do coeficiente linear de pearson. *Ciência & Engenharia*, 15(1/2):45–53.

Löwe, B., Kroenke, K., and Gräfe, K. (2005). Detecting and monitoring depression with a two-item questionnaire (phq-2). *Journal of psychosomatic research*, 58(2):163–171.

Monard, M. C. and Baranauskas, J. A. (2003). Indução de regras e árvores de decisão. *Sistemas Inteligentes. Rezende, SO Editora Manole Ltda*, pages 115–140.

Morettin, P. A. and BUSSAB, W. O. (2017). *Estatística básica*. Editora Saraiva.

Murray, J. F., Hughes, G. F., and Kreutz-Delgado, K. (2005). Machine learning methods for predicting failures in hard drives: A multiple-instance application. *Journal of Machine Learning Research*, 6(May):783–816.

Niaura, R., Britt, D. M., Borrelli, B., Shadel, W. G., Abrams, D. B., and Goldstein, M. G. (1999). History and symptoms of depression among smokers during a self-initiated quit attempt. *Nicotine & Tobacco Research*, 1(3):251–257.

Pasquali, L. (2009). Psychometrics. *Revista da Escola de Enfermagem da USP*, 43(SPE):992–999.

Patel, M. J., Khalaf, A., and Aizenstein, H. J. (2016). Studying depression using imaging and machine learning methods. *NeuroImage: Clinical*, 10:115–123.

Pinto Gomide, H., Rodrigues Teixeira de Carvalho, C., Lovisi Menezes, M., Gazolla de Oliveira, I., Furtado de Mendonça, G., Duque de Albuquerque Júnior, R., Munck Machado, N., Russi Ervilha, R., Costa Rizuti da Rocha, T., Soares Bernardino, H., et al. (2018). Depression among smokers of a web-based intervention to quit smoking: a cross-sectional study. *Salud Mental*, 40(6):271–277.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richardson, L. P., Rockhill, C., Russo, J. E., Grossman, D. C., Richards, J., McCarty, C., McCauley, E., and Katon, W. (2010). Evaluation of the phq-2 as a brief screen for detecting major depression among adolescents. *Pediatrics*, 125(5):e1097.

Rossi, R. A. and Ahmed, N. K. (2015). The network data repository with interactive graph analytics and visualization. In *AAAI*.

Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

SOUZA, S., B. H., and L, G. (2016). ProgramaÇÃo genÉtica para a previsÃo de propriedadesmecÂnicas de concretos de agregados leves. In *XII SIMMEC, Simpósio de Mecânica Computacional*. SIMMEC.

Spitzer, R., Williams, J., and Kroenke, K. (1999). Instruction manual: Instructions for patient health questionnaire (phq) and gad-7 measures.

Staples, L. G., Dear, B. F., Gandy, M., Fogliati, V., Fogliati, R., Karin, E., Nielssen, O., and Titov, N. (2019). Psychometric properties and clinical utility of brief measures of depression, anxiety, and general distress: The phq-2, gad-2, and k-6. *General hospital psychiatry*, 56:13–18.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89.

Tanner, L., Schreiber, M., Low, J. G., Ong, A., Tolfvenstam, T., Lai, Y. L., Ng, L. C., Leo, Y. S., Puong, L. T., Vasudevan, S. G., et al. (2008). Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS neglected tropical diseases*, 2(3):e196.

Tchounga, B. K., Inwoley, A., Coffie, P. A., Minta, D., Messou, E., Bado, G., Minga, A., Hawerlander, D., Kane, C., Eholie, S. P., et al. (2014). Re-testing and misclassification of hiv-2 and hiv-1&2 dually reactive patients among the hiv-2 cohort of the west african database to evaluate aids collaboration. *Journal of the International AIDS Society*, 17(1):19064.

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., and Sunderland, M. (2011). Psychometric comparison of the phq-9 and bdi-ii for measuring response during treatment of depression. *Cognitive Behaviour Therapy*, 40(2):126–136.

Veiga, R. V., de Freitas, J. M., Bernardino, H. S., Barbosa, H. J., and Alcântara-Neves, N. M. (2015). Using grammar-based genetic programming to determine characteristics of multiple infections and environmental factors in the development of allergies and asthma. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, pages 1604–1611. IEEE.

WHO (2017). *WHO REPORT ON THE GLOBAL TOBACCO EPIDEMIC: Monitoring tobacco use and prevention policies*. WHO.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20(7):557–585.

Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., and Sahli, H. (2016). Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 89–96. ACM.

Zeng, X. and Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12.

A Questionários psicométricos

QUESTIONÁRIO SOBRE A SAÚDE DO PACIENTE- 9 (P H Q - 9)

Durante os <u>últimos 14 dias</u>, em quantos foi afectado/a por algum dos seguintes problemas? *(Utilize"✔" para indicar a sua resposta)*	**Nunca**	**Em vários dias**	**Em mais de metade do número de dias**	**Em quase todos os dias**
1. Tive pouco interesse ou prazer em fazer coisas	0	1	2	3
2. Senti desânimo, desalento ou falta de esperança	0	1	2	3
3. Tive dificuldade em adormecer ou em dormir sem interrupções, ou dormi demais	0	1	2	3
4. Senti cansaço ou falta de energia	0	1	2	3
5. Tive falta ou excesso de apetite	0	1	2	3
6. Senti que não gosto de mim próprio/a — ou que sou um(a) falhado/a ou me desiludi a mim próprio/a ou à minha família	0	1	2	3
7. Tive dificuldade em concentrar-me nas coisas, como ao ler o jornal ou ver televisão	0	1	2	3
8. Movimentei-me ou falei tão lentamente que outras pessoas poderão ter notado. Ou o oposto: estive agitado/a a ponto de andar de um lado para o outro muito mais do que é habitual	0	1	2	3
9. Pensei que seria melhor estar morto/a, ou em magoar-me a mim próprio/a de alguma forma	0	1	2	3

FOR OFFICE CODING *0* **+** _____ **+** _____ **+** _____

=Total Score: _____

Se indicou <u>alguns</u> problemas, até que ponto é que eles <u>dificultaram</u> o seu trabalho, o cuidar da casa ou o lidar com outras pessoas?

Não dificultaram	**Dificultaram um pouco**	**Dificultaram muito**	**Dificultaram extremamente**
☐	☐	☐	☐

Desenvolvido por Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke e colegas, com uma bolsa de estudos da Pfizer Inc.

TESTE DE FAGERSTRÖM

	pontos	soma
Quanto tempo depois de acordar você fuma o primeiro cigarro?		
após 60 minutos	0	
entre 31 e 60 minutos	1	
entre seis e 30 mintuos	2	
nos primeiros cinco minutos	3	
Você encontra dificuldades em evitar de fumar em locais proibidos, como por exemplo: igrejas, local de trabalho, cinemas, shoppings, etc?		
não	0	
sim	1	
Qual o cigarro mais difícil de largar de fumar?		
qualquer outro	0	
o primeiro da manhã	1	
Quantos cigarros você fuma por dia?		
menos de 10 cigarros	0	
entre 11 e 20 cigarros	1	
entre 21 e 30 cigarros	2	
mais de 30 cigarros	3	
Você fuma mais freqüentemente nas primeiras horas do dia do que durante o resto do dia?		
não	0	
sim	1	
Você fuma mesmo estando doente ao ponto de ficar acamado na maior parte do dia?		
não	0	
sim	1	

Pontuação:
0 a 4 – dependência leve; 5 a7 – dependência moderada e 8 a 10 – dependência grave

II Consenso Brasileiro de DPOC 2004 (modificado de Fagestrom K 1989)

Questionário AUDIT

1. Com que frequência consome bebidas que contêm álcool? *[Escreva o número que melhor corresponde à sua situação.]*
0 = nunca
1 = uma vez por mês ou menos
2 = duas a quatro vezes por mês
3 = duas a três vezes por semanas
4 = quatro ou mais vezes por semana

2. Quando bebe, quantas bebidas contendo álcool consome num dia normal?
0 = uma ou duas
1 = três ou quatro
2 = cinco ou seis
3 = de sete a nove
4 = dez ou mais

3. Com que frequência consome seis bebidas ou mais numa única ocasião?
0 = nunca
1 = menos de um vez por mês
2 = pelo menos uma vez por mês
3 = pelo menos uma vez por semana
4 = diariamente ou quase diariamente

4. Nos últimos 12 meses, com que frequência se apercebeu de que não conseguia parar de beber depois de começar?
0 = nunca
1 = menos de um vez por mês
2 = pelo menos uma vez por mês
3 = pelo menos uma vez por semana
4 = diariamente ou quase diariamente

5. Nos últimos 12 meses, com que frequência não conseguiu cumprir as tarefas que habitualmente lhe exigem por ter bebido?
0 = nunca
1 = menos de um vez por mês
2 = pelo menos uma vez por mês
3 = pelo menos uma vez por semana
4 = diariamente ou quase diariamente

6. Nos últimos 12 meses, com que **frequência precisou de beber** logo **de manhã** para "curar" uma ressaca?
0 = nunca
1 = menos de um vez por mês
2 = pelo menos uma vez por mês
3 = pelo menos uma vez por semana
4 = diariamente ou quase diariamente

7. Nos últimos 12 meses, com que **frequência** teve **sentimentos de culpa ou** de **remorsos** por **ter bebido**?
0 = nunca
1 = menos de um vez por mês
2 = pelo menos uma vez por mês
3 = pelo menos uma vez por semana
4 = diariamente ou quase diariamente

8. Nos últimos 12 meses, com que **frequência <u>não</u> se lembrou** do que **aconteceu na noite anterior por** causa de **ter bebido**?
0 = nunca
1 = menos de um vez por mês
2 = pelo menos uma vez por mês
3 = pelo menos uma vez por semana
4 = diariamente ou quase diariamente

9. Já alguma vez ficou ferido ou **ficou alguém ferido** por **<u>você</u> ter bebido**?
0 = não
1 = sim, mas não <u>nos últimos 12 meses</u>
2 = sim, aconteceu <u>nos últimos 12 meses</u>

10. Já alguma vez um **familiar**, **amigo**, **médico ou profissional de saúde** manifestou **preocupação** pelo **seu consumo de álcool** ou sugeriu que **deixasse de beber**?
0 = não
1 = sim, mas não <u>nos últimos 12 meses</u>
2 = sim, aconteceu <u>nos últimos 12 meses</u>

B Variáveis do Banco de Dados

Tabela B.1: Variáveis do sistema.

Variável	**Tipo**	**Valores**	**Valores Faltantes**
alcohol_use_1	Inteiro	[0,3]	242
alcohol_use_2	Inteiro	[0,3]	242
alcohol_use_3	Inteiro	[0,3]	242
dependency_level_1	Binária	[0,3]	242
dependency_level_2	Binária	0 ou 1	242
dependency_level_3	Binária	0 ou 1	242
dependency_level_4	Inteiro	[0,80]	242
dependency_level_5	Binária	0 ou 1	242
dependency_level_6	Binária	0 ou 1	242
education	categórica	1 = Ensino fundamental 2 = Ensino Médio 3 = Ensino Superior 4 = Pós Graduação	242
employed	Lógico	0 = Não 1 = Sim	242
problems_1	Inteiro	[0,3]	242
problems_2	Inteiro	[0,3]	242
ftnd_1	Inteiro	[0,3]	0
ftnd_2	Binária	0 ou 1	0
ftnd_3	Binária	0 ou 1	0
ftnd_4	Inteiro	[0,80]	0
ftnd_5	Binária	0 ou 1	0
ftnd_6	Binária	0 ou 1	0

Tabela B.1: Variáveis do sistema.

Variável	**Tipo**	**Valores**	**Valores Faltantes**
pnad_a	Lógico	0 = Não 1 = Sim	0
pnad_b	Lógico	0 = Não 1 = Sim	0
pnad_d	Lógico	0 = Não 1 = Sim	0
pnad_e	Lógico	0 = Não 1 = Sim	0
pnad_f	Lógico	0 = Não 1 = Sim	0
pnad_g	Lógico	0 = Não 1 = Sim	0
pnad_h	Lógico	0 = Não 1 = Sim	0
ladder	Inteiro	[1,10]	8
enfrentar_fissura_comer	Lógico	0 = Não 1 = Sim	0
enfrentar_fissura_ler_razoes	Lógico	0 = Não 1 = Sim	0
tempo_para_parar_dias	Inteiro	0 = Já parou 1 = Até 7 dias 2 = Mais de 7 dias	0
tentou_parar	Lógico	0 = Não 1 = Sim	0
age	discreta	[18,70]	0
gender	categórica	0 = Masculino 1 = Feminino	0

Tabela B.1: Variáveis do sistema.

Variável	**Tipo**	**Valores**	**Valores Faltantes**
phq_1	discreta	[0,3]	0
phq_2	discreta	[0,3]	0
phq_3	discreta	[0,3]	0
phq_4	discreta	[0,3]	0
phq_5	discreta	[0,3]	0
phq_6	discreta	[0,3]	0
phq_7	discreta	[0,3]	0
phq_8	discreta	[0,3]	0
phq_9	discreta	[0,3]	0