Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Estudo da prevalência de transtornos mentais comuns via métodos de classificação

Giulia Molina Martinez

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Estudo da prevalência de transtornos mentais comuns via métodos de classificação

Giulia Molina Martinez
Orientadora: Profª Drª Andressa Cerqueira

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Fevereiro de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Study of the prevalence of common mental disorders through classification methods

Giulia Molina Martinez
Advisor: Profª Drª Andressa Cerqueira

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
February 2024

Giulia Molina Martinez

Estudo da prevalência de transtornos mentais comuns via métodos de classificação

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Giulia Molina Martinez e aprovado pela banca examinadora.

Aprovado em 26 de janeiro de 2024

Banca Examinadora:

- Profª. Drª. Andressa Cerqueira
- Profª. Drª. Teresa Cristina Martins Dias
- Prof. Ms. Pedro Ferreira Filho

Resumo

O presente Trabalho de Graduação consiste no estudo da associação de determinadas variáveis com o diagnóstico de transtornos mentais comuns (TMC). Para isso, estudamos métodos de classificação, sendo eles, árvores de classificação, regressão logística e florestas aleatórias, antecedidos de uma definição estatística e revisão metodológica dos conceitos a serem utilizados e do pré-processamento dos dados. Todas as aplicações computacionais foram feitas através da linguagem de programação `R` (R Development Core Team, 2023).

Palavras-chave: *árvores de classificação, classificação, florestas aleatórias, regressão logística, transtornos mentais comuns.*

Abstract

The present undergraduate thesis focuses on the study of the association between specific variables and the diagnosis of common mental disorders. To achieve this goal, we examined classification methods, namely decision trees, logistic regression, and random forests. This investigation is preceded by a statistical definition and methodological review of the concepts to be employed, as well as data preprocessing. All computational applications were carried out using the programming language `R` (R Development Core Team, 2023).

Keywords: *classification trees, classification, common mental disorders, logistic regression, random forests.*

Sumário

Capítulo 1

Introdução

A estatística é uma ciência que lida com a coleta, análise, interpretação, apresentação e organização de dados. Ela desempenha um papel fundamental em diversos campos como, exemplo, no campo da saúde. Sendo assim, este trabalho tem por interesse utilizar algoritmos de classificação binária para identificar quais covariáveis tem maior poder explicativo para o diagnóstico de transtornos mentais comuns (TMC). A presença de TMC foi definida a partir da variável SQR obtida por meio de um questionário de 20 perguntas com respostas dicotômicas (sim/não), denominado SQR20 (*Self-Reporting Questionnaire*). Dessa forma, as respostas positivas são sumarizadas em um escore que, a partir de um ponto de corte definido, indica a presença de TMC.

Nesse trabalho o SQR20 foi categorizado em dois níveis, sendo eles, 0 (ausência de TMC) e 1 (presença de TMC). Essa categorização foi feita a partir de pontos de corte específicos, definidos para adultos/adolescentes e idosos, definidos da seguinte forma:

Adultos/Adolescentes (de Jesus Mari e Williams, 1986):

- Masculino: 6 ou mais respostas "sim"= presença de TMC;
- Feminino: 8 ou mais respostas "sim"= presença de TMC;

Idosos ($\geq$ 60 anos) (Scazufca *et al.*, 2009):

- Masculino e Feminino: 5 ou mais respostas "sim"= presença de TMC.

Os transtornos mentais são caracterizados por alterações do modo de pensar e do humor, ou por comportamentos associados com angústia pessoal e/ou deterioração do funcionamento (Araújo e Neto, 2014). Transtornos mentais comuns são descritos por pensamentos depressivos, ansiedade, irritabilidade, insônia, fadiga, dificuldade de memória e

concentração e queixas somáticas. Vale ressaltar que é fundamental o diagnóstico precoce de TMC para garantir que as pessoas recebam o suporte adequado, reduzam o sofrimento, previnam complicações e melhorem sua qualidade de vida.

No artigo Bastos *et al.* (2020), os autores apresentam uma aplicação utilizando regressão logística em dois diferentes modelos, a fim de investigar a associação entre a presença de transtornos mentais comuns e a aderência ao padrão alimentar mediterrâneo tradicional e ao padrão alimentar mediterrâneo brasileiro (incluí alimentos com características não mediterrânicas).

Em outra análise, Rahman *et al.* (2015) apresentam uma análise do efeito da carga de trabalho mental sobre o estresse mental humano por meio da seleção de características do sinal de eletroencefalograma (EEG) e classificação para reconhecer o estresse mental. Vários sinais de EEG foram coletados e analisados utilizando análises de espectro e características foram extraídas usando o classificador k-vizinhos Mais Próximos (KNN).

O objetivo principal deste trabalho está em estudar a prevalência de transtornos mentais comuns (TMC) via métodos de classificação binária e comparar estes por meio de métricas avaliativas. Como objetivo secundário temos estudar e revisar os métodos de classificação: árvores de classificação, regressão logística e florestas aleatórias e formalizar de forma técnica os algoritmos desenvolvidos para cada um dos métodos.

Este trabalho está organizado como segue. No Capítulo 1, é apresentada a introdução do trabalho. No Capítulo 2, revisamos os métodos de classificação binária que foram aplicados, sendo eles: Árvores de Classificação, Florestas Aleatórias e Regressão Logística, seguidos da apresentação das medidas de desempenho e pontos de corte que serão utilizadas para compararmos a performance de cada um dos algoritmos e descrição do banco de dados. No Capítulo 3, realizamos uma análise descritiva e exploratória das covariáveis utilizadas em relação a variável resposta e apresentamos os resultados obtidos em cada método de classificação binária aplicado. Por fim, no Capítulo 4 finalizamos com as considerações finais.

Capítulo 2

Materiais e Métodos

Problemas de classificação são similares aos problemas de predição em regressão, uma vez que, em ambos os cenários consideramos uma amostra com observações independentes a fim de construir uma função $g(x)$ que seja capaz de fornecer uma boa previsão para futuras observações. A distinção entre um problema de classificação e um problema de regressão reside no fato de que, no primeiro, a variável resposta não é uma variável quantitativa, mas sim qualitativa (Izbicki e dos Santos, 2020).

Neste capítulo apresentamos uma revisão dos métodos de classificação binária utilizados para estudar a prevalência de transtornos mentais comuns (TMC), sendo eles, árvores de classificação, regressão logística e florestas aleatórias, também abordamos as métricas avaliativas que foram usadas para comparar os resultados de cada um dos métodos aplicados e fizemos a descrição do banco de dados utilizado.

Inicialmente, revisamos o conceito de *data-splitting* que foi abordado nos métodos de classificação. *Data-splitting* refere-se à prática de dividir o conjunto de dados disponível em subconjuntos distintos, essa técnica é usada para evitar o super-ajuste (*overfitting*) e garantir que as métricas de desempenho obtidas sejam representativas do verdadeiro poder de generalização do modelo. Geralmente, o processo de *data-splitting* envolve a criação de um conjunto de treinamento, usado para estimar a função de predição e um conjunto de validação, usado para a avaliar o poder preditivo dos modelos selecionados.

2.1 Árvores de Classificação

Árvores de classificação são similares à árvores de regressão, no entanto tem o objetivo específico de realizar a classificação. Sendo assim, para um melhor entendimento do

método, iniciamos revisando o conceito de árvores de regressão.

2.1.1 Árvores de regressão

Árvores de regressão fazem parte dos métodos não paramétricos, visto que, não requer que a distribuição da população seja caracterizada por certos parâmetros. Devido a isso, é considerado um método mais flexível, pois requer suposições geralmente mais fracas sobre a população a partir da qual os dados são obtidos. Outra característica importante é a fácil interpretação e aplicação do método, em que a função de regressão estimada é sempre constante por partes.

A ideia principal da criação de uma árvore está em dividir o espaço das covariáveis em uma partição $R_1, \ldots, R_k$. Considere um sequência $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, em que y_i é a variável resposta e $\mathbf{x}_i$ é um vetor contendo as d covariáveis para a i-ésima observação. A predição para a resposta Y de uma observação com covariáveis $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} \in R_k$, é feita da seguinte forma (Izbicki e dos Santos, 2020):

$$g(\mathbf{x}) = \frac{1}{|\{i : \mathbf{x}_i \in R_k\}|} \sum_{i:\mathbf{x}_i \in R_k} y_i,$$

isto é, a predição da resposta de $\mathbf{x}$ (folhas) é feita a partir da média dos valores da variável resposta das amostras do conjunto de treinamento pertencentes àquela mesma região.

A partir disso, a criação de uma árvore de regressão é feita em duas etapas:

1. criação de uma árvore completa e complexa;
2. poda.

Na etapa 1 buscamos criar uma árvores com partições homogêneas, a fim de avaliar o quão "pura" uma árvore T é usamos o erro o quadrático médio,

$$P(T) = \sum_{R} \sum_{i:\mathbf{x}_i \in R} \frac{(y_i - \hat{y}_R)^2}{n}$$

em que, $\hat{y}_R$ é o valor predito para uma observação pertencente à região R.

O procedimento de construção da árvore é feito através de parcionamentos recursivos no espaço das covariáveis até obter uma árvore com poucas observações por folhas. A

árvore produzida na etapa 1 apresenta bons resultados para o conjunto de treinamento, mas é provável que ocorra super-ajuste (*overfitting*), o que gera uma predição ruim para novas observações, sendo assim, seguimos para a etapa de poda.

Na etapa 2 o objetivo é tornar a árvore de regressão menor e menos complexa, diminuindo assim a variância desse estimador. Para isso, é retirado um nó por vez e observa-se a variação do erro de predição no conjunto de validação. A partir disso, decide-se quais nós permanecerão na árvore.

Na Figura 2.1 ilustramos a estrutura de uma árvore de regressão.

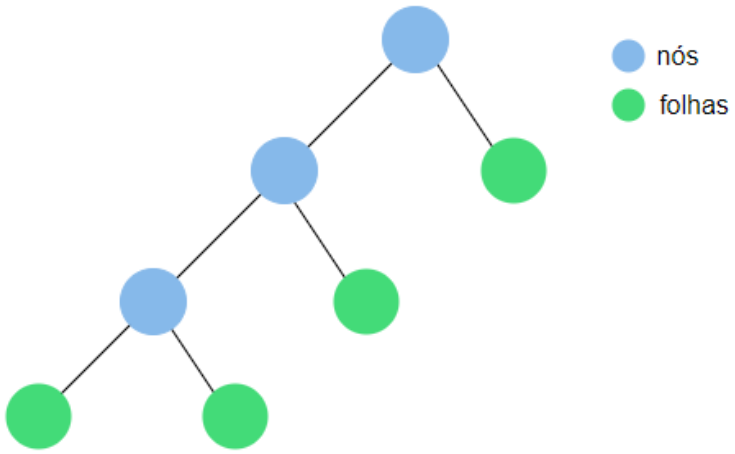


Figura 2.1: Exemplo de estrutura de uma árvore de regressão.

Utilizamos uma árvore para prever novas observações da seguinte forma: começamos pelo topo, verificamos se a condição desse nó é satisfeita, se sim, seguimos a esquerda, caso contrário, seguimos a direita. Dessa forma, seguimos até que uma folha seja atingida.

2.1.2 Árvores de classificação

Agora, após revisar os conceitos básicos de árvores de regressão, apontamos duas diferenças em relação às árvores de classificação, na qual assume-se que a variável resposta é categórica, sendo C o conjunto de possíveis categorias.

(i) A predição para a resposta Y de uma observação com covariáveis $\mathbf{x}$ não é mais dada pela média das observações mas sim pela moda, ou seja,

$$g(\mathbf{x}) = \text{moda}\{y_i : \mathbf{x}_i \in R_k\}, \quad \text{se} \quad \mathbf{x} \in R_k.$$

(ii) O critério utilizado na etapa 1 é diferente, agora, usamos o índice de Gini (Hastie *et al.*, 2009), uma vez que esse é sensível à mudanças nas proporções de cada categoria nos nós.

$$\sum_{R}\sum_{c\in C}\hat{p}_{R,c}(1-\hat{p}_{R,c}),$$

em que, R representa uma das regiões induzidas pela árvore e $\hat{p}_{R,c}$ é a proporção de observações classificadas como sendo da categoria c entre as que caem na região R.

Como citado em Hastie *et al.* (2009), outros critérios podem ser utilizados para substituir o erro quadrático médio, sendo eles, *misclassification error* e *deviance*. Nesse trabalho, optamos por utilizar o índice de Gini, em que um pequeno valor indica que um nó contém predominantemente observações de uma única classe.

Vale ressaltar que, a maneira como árvores são construídas faz com que covariáveis irrelevantes sejam descartadas. Sendo assim, a seleção de variáveis é feita automaticamente e, também, não é necessário incluir termos de interação adicionais, uma vez que, a árvore lida naturalmente com interações entre variáveis.

2.2 Florestas Aleatórias

Como visto anteriormente, as árvores de classificação possuem alta interpretabilidade porém costumam apresentar baixo poder preditivo quando comparadas aos demais estimadores (Izbicki e dos Santos, 2020). Sendo assim, a metodologia de florestas aleatórias (Breiman, 2001) combina diversas árvores diferentes e não correlacionadas para a tomada de decisão a fim de obter um classificador com maior poder preditivo.

O método de florestas aleatórias consistem em criar B árvores distintas, para isso utiliza B amostras bootstrap (da Silva Filho, 2010) da amostra original, em que cada nó só é permitido que seja escolhida uma dentre as $m < d$ covariáveis. O subconjunto de covariáveis m são escolhidas aleatoriamente dentre as covariáveis originais e, a cada nó criado, um novo subconjunto de covariáveis é sorteado (Izbicki e dos Santos, 2020).

A combinação das diversas árvores de classificação é realizada por meio da aplicação da função:

$$g(\mathbf{x}) = \text{moda}\{g^b(\mathbf{x}), b = 1, \ldots, B\}$$

ou seja, cada árvore construída classifica uma observação que possui covariáveis $\mathbf{x}$, e a previsão final é determinada pela categoria mais frequentemente predita entre todas as árvores.

Vale ressaltar que, o valor de m pode ser escolhido via validação cruzada. Estudos já realizados indicam, em linhas gerais, que quando o número de covariáveis (m) é aproximadamente um terço do número total de covariáveis (d), o desempenho do modelo tende a ser satisfatório.

2.3 Regressão logística

O método de regressão logística, diferentemente dos métodos vistos anteriormente, é um método paramétrico, ou seja, a estimativa da função de regressão necessariamente pertence a um espaço de funções que podem ser parametrizadas por um número finito de parâmetros. Sendo assim, o objetivo é encontrar os valores desses parâmetros que melhor se ajustam aos dados.

Nos modelos de regressão logística binária, a variável resposta Y é dicotômica, assumindo apenas os valores 0 e 1. Dessa forma, a regressão logística modela através de uma variável aleatória com distribuição de bernoulli a probabilidade de Y pertencer a uma determinada categoria, garantindo, ao mesmo tempo, que as probabilidades das classes variem entre 0 e 1.

Como visto no livro James *et al.* (2013), sempre que uma linha reta é ajustada a uma resposta binária codificada como 0 ou 1, em princípio podemos prever valores menores que 0 ou maiores que 1 (a menos que o intervalos de $\mathbf{X}$ seja limitado). A fim de evitar isso, modelamos $\mathbb{P}(Y = 1|\mathbf{x})$ usando uma função com saídas entre 0 e 1 para qualquer $\mathbf{x}$. Na regressão logística, utilizamos a função logística,

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}, \quad i = 1, \ldots, d,$$

em que $\mathbf{x} = (1, x_1, \ldots, x_d)$ denota o vetor com a constante 1 e os valores observados das covariáveis, β_0 é uma constante e β_i são os d parâmetros de regressão. Esse modelo fornece

a probabilidade do indivíduo apresentar a resposta de interesse, sendo assim,

$$1 - \mathbb{P}(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}},$$

fornece a probabilidade do indivíduo não apresentar a resposta de interesse.

A transformação em $\mathbb{P}(Y = 1|\mathbf{x})$ definida pelo logaritmo neperiano da razão entre $\mathbb{P}(Y = 1|\mathbf{x})$ e $1 - \mathbb{P}(Y = 1|\mathbf{x})$, chamada de logito, fornece um modelo linear (Giolo, 2021):

$$\ln\left[\frac{\mathbb{P}(Y = 1|\mathbf{x})}{1 - \mathbb{P}(Y = 1|\mathbf{x})}\right] = \beta_0 + \sum_{i=1}^{d} \beta_i x_i = \boldsymbol{\beta'}\mathbf{x}.$$

A razão entre $\mathbb{P}(Y = 1|\mathbf{x})$ e $1 - \mathbb{P}(Y = 1|\mathbf{x})$ define uma chance (*odds*), portanto o logito é o logaritmo de uma chance, dessa forma,

$$\text{chance} = \frac{\mathbb{P}(Y = 1|\mathbf{x})}{1 - \mathbb{P}(Y = 1|\mathbf{x})} = e^{\boldsymbol{\beta'}\mathbf{x}}.$$

Para estimar os coeficientes de uma regressão logística, utilizamos o método de máxima verossimilhança (Izbicki e dos Santos, 2020). Dada uma amostra i.i.d, independente e identicamente distribuída, $(\mathbf{X}_1, Y_1)$, ..., $(\mathbf{X}_n, Y_n)$, a função de verossimilhança condicional é:

$$\begin{aligned} L(y; (\boldsymbol{x}, \boldsymbol{\beta})) &= \prod_{k=1}^{n} (\mathbb{P}(Y_k = 1|\boldsymbol{x}_k, \boldsymbol{\beta}))^{y_k} (1 - \mathbb{P}(Y_k = 1|\boldsymbol{x}_k, \boldsymbol{\beta}))^{1-y_k} \\ &= \prod_{k=1}^{n} \left(\frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{k,i}}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{k,i}}}\right)^{y_k} \left(\frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{k,i}}}\right)^{1-y_k}. \end{aligned}$$

Utilizando métodos numéricos para maximizar $L(y; (\boldsymbol{x}, \boldsymbol{\beta}))$, obtém-se os estimadores $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_d$, respectivamente, dos parâmetros β_0, β_1, ..., β_d.

A regressão logística apesar de ter suposições mais fortes em relação aos métodos não paramétricos, como a de linearidade entre as covariáveis e a variável resposta, apresenta vantagens em relação a sua eficiência computacional e fácil interpretabilidade pois fornece coeficientes para cada variável independente, permitindo uma interpretação direta do impacto de cada covariável na variável resposta.

2.4 Medidas de Desempenho

As medidas de desempenho de um modelo são utilizadas para avaliar o quão bem o modelo está realizando suas previsões ou classificações em relação aos dados reais. Sendo assim, algumas medidas de desempenho definidas a partir de uma matriz de confusão a fim de comparar os modelos obtidos através dos diferentes métodos de classificação aplicados, sendo elas: Acurácia, Sensibilidade, Especificidade, Valor preditivo positivo, Valor preditivo negativo e Estatística F1.

Uma matriz de confusão é representada por uma tabela usada para descrever o desempenho de um modelo de classificação, ou seja, é uma representação tabular das previsões feitas pelo modelo em comparação com as classes reais dos dados. Na Tabela 2.1 podemos observar como uma matriz de confusão é construída (Izbicki e dos Santos, 2020).

Tabela 2.1: Matriz de Confusão.

Valor Predito	**Valor Verdadeiro**	
	$Y = 0$	$Y = 1$
$Y = 0$	VN	FN
$Y = 1$	FP	VP

Com base na Tabela 2.1, temos que VN (Verdadeiros Negativos) representa a classificação correta da classe negativo, FN (Falsos Negativos) aborda o erro tipo II, ou seja, o modelo previu a classe negativo quando o valor real era classe positivo, FP (Falsos Positivos) aborda o erro tipo I, ou seja, o modelo previu a classe positivo quando o valor verdadeiro era classe negativo e VP (Verdadeiros Positivos) representa a classificação correta da classe positivo.

A partir da Tabela 2.1, podemos definir as diferentes medidas de desempenho, dadas por:

- **Acurácia**: A $= \frac{VN+VP}{VN+FN+FP+VP}$;
- **Sensibilidade ou *Recall***: S $= \frac{VP}{VP+FN}$;
- **Especificidade**: E $= \frac{VN}{VN+FP}$;
- **Valor preditivo positivo ou Precisão**: VPP $= \frac{VP}{VP+FP}$;
- **Valor preditivo negativo**: VPN $= \frac{VN}{VN+FN}$;
- **Estatística F1**: F1 $= \frac{2}{\frac{1}{S}+\frac{1}{VPP}}$ (média harmônica entre S e VPP).

A interpretação de cada uma das medidas de desempenho citadas é descrita a seguir:

- **Acurácia**: avalia a proporção de previsões corretas feitas por um modelo em relação ao total de previsões realizadas, ou seja, mede a capacidade do modelo de classificar corretamente as observações em todas as classes;
- **Sensibilidade ou *Recall***: proporção de obervações positivas corretamente identificadas pelo modelo em relação ao número total de obervações verdadeiramente positivas, em outras palavras, mede a capacidade do modelo em capturar corretamente os casos positivos;
- **Especificidade**: proporção de obervações negativas corretamente identificadas pelo modelo em relação ao número total de obervações verdadeiramente negativas, em outras palavras, avalia a habilidade do modelo em identificar casos negativos de forma precisa;
- **Valor preditivo positivo ou Precisão**: proporção de obervações verdadeiramente positivas em relação ao total de obervações classificadas como positivas pelo modelo, sendo assim, avalia a precisão das previsões positivas do modelo;
- **Valor preditivo negativo**: proporção de obervações verdadeiramente negativas em relação ao total de obervações classificadas como negativas pelo modelo, sendo assim, avalia a precisão das previsões negativas do modelo;
- **Estatística F1**: fornece uma medida única de desempenho que leva em consideração tanto os verdadeiros positivos quanto os falsos positivos e falsos negativos.

Vale destacar que, as estatísticas calculadas com base na Tabela 2.1 são estimativas populacionais. Dessa forma, é necessário calcular os valores de VP, FN, VN e FP aplicando uma amostra de teste ou validação para evitar o super-ajuste (*overfitting*).

2.5 Dados Desbalanceados e Outros Cortes

Dados desbalanceados referem-se a conjuntos de dados em que as classes que se deseja prever contêm um número desigual de elementos. Essa situação pode apresentar desafios para os modelos de aprendizado de máquina, pois eles podem ter uma tendência a favorecer a classe majoritária, resultando em um desempenho inferior para a classe minoritária.

Quando se lida com dados desbalanceados, uma estratégia comum é ajustar os pontos de corte utilizados para tomar decisões de classificação. O ponto de corte padrão geralmente é definido em $0,5$ para problemas binários, o que significa que, se a probabilidade de pertencer à classe 1 for maior que $0,5$, o modelo prevê essa classe.

Uma abordagem comum para resolver essa questão envolve a busca por valores de ponto de corte K distintos de $0,5$ nos modelos de classificação baseados em probabilidades. Com isso, procuramos:

$$g(\boldsymbol{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq K).$$

A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica usada para avaliar o desempenho de um modelo de classificação binária em diferentes pontos de corte para tomada de decisão. A curva ROC mostra a relação entre a Taxa de Verdadeiros Positivos (Sensibilidade) e a Taxa de Falsos Positivos (1 - Especificidade) em vários pontos de corte.

Cada ponto na curva ROC representa o desempenho do modelo em um determinado ponto de corte. Quanto mais próxima a curva estiver do canto superior esquerdo do gráfico, melhor será o desempenho do modelo, indicando uma alta sensibilidade e uma baixa taxa de falsos positivos. Além disso, temos a área sob a curva ROC (AUC-ROC) que é uma métrica comum para resumir o desempenho global do modelo, em que um valor igual a 1 indica um desempenho perfeito.

A partir da curva ROC (Fawcett, 2006) temos diversas abordagens para escolher o melhor ponto de corte K, sendo duas delas, a média geométrica (Média-G) e a estatística de Youden (J).

- Média Geométrica: é uma métrica de avaliação de desempenho em classificação desbalanceada, buscando um equilíbrio entre sensibilidade e especificidade por meio da média geométrica dessas métricas. A fórmula matemática para a G-Média é expressa por:

$$\text{Média-G} = \sqrt{\text{Sensibilidade} \cdot \text{Especificidade}}.$$

- Estatística de Youden: é uma medida que busca otimizar simultaneamente a sen-

sibilidade e a especificidade, sendo utilizada para encontrar um ponto de corte que equilibre essas métricas. Sua expressão matemática é dada por:

$$\text{J} = \text{Sensibilidade} + \text{Especificidade} - 1.$$

2.6 Descrição do Banco de Dados

O conjunto de dados utilizado é proveniente do projeto temático "Estilo de vida, marcadores bioquímicos e genéticos como fatores de risco cardiometabólico: inquérito de saúde na cidade de São Paulo" processo FAPESP 17/05125-7. Este contém a informação de 3722 indivíduos e, com o objetivo de estudar a prevalência de transtornos mentais comuns (TMC) via métodos de classificação binária, selecionamos 23 covariáveis, com o auxílio da pesquisadora, das quais 7 são quantitativas e 16 qualitativas.

A variável resposta escolhida (SQR_cat) foi construída através do SQR20 (*Self-Reporting Questionnaire*) e categorizada em dois níveis, da seguinte forma:

$$Y = \begin{cases} 0, \text{ ausência de TMC;} \\ 1, \text{ presença de TMC.} \end{cases}$$

Para auxiliar a interpretação, são descritas as covariáveis utilizadas no estudo, divididas em blocos, da mesma forma que foram definidas no dicionário do banco de dados disponibilizado:

1. **Variáveis de identificação:**

   - Idade: idade do indivíduo;
   - Sexo:

$$\text{Sexo} = \begin{cases} 1 \rightarrow \text{ Feminino;} \\ 2 \rightarrow \text{ Masculino.} \end{cases}$$

2. **Qualidade de vida:**

- F101: indica como o indivíduo diria que está seu estado de saúde.

$$\text{F101} = \begin{cases} 1 \rightarrow \text{ Excelente/Muito boa;} \\ 2 \rightarrow \text{ Boa;} \\ 3 \rightarrow \text{Regular;} \\ 4 \rightarrow \text{Ruim;} \\ 5 \rightarrow \text{Muito ruim;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

3. **Atividade física:**

- K201a: indica se atualmente o indivíduo trabalha ou faz trabalho voluntário fora de sua casa.

$$\text{K201a} = \begin{cases} 1 \rightarrow \text{ Não;} \\ 2 \rightarrow \text{ Sim;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

- K205a_hrs: indica o tempo total (em horas) que o indivíduo gasta sentado durante um dia de semana.
- K205b_hrs: indica o tempo total (em horas) que o indivíduo gasta sentado durante um dia de final de semana.
- K206_hrs: indica o tempo total (em horas) que o indivíduo gasta assistindo TV durante um dia de semana.
- K207_hrs: indica o tempo total (em horas) que o indivíduo gasta assistindo TV durante um dia de final de semana.
- K208_hrs: indica o tempo total (em horas) que o indivíduo gasta no computador durante um dia de semana.
- K209_hrs: indica o tempo total (em horas) que o indivíduo gasta no computador durante um dia de final de semana.
- K210: indica se o indivíduo pratica regularmente, pelo menos uma vez por

semana, algum tipo de exercício físico ou esporte.

$$\text{K210} = \begin{cases} 1 \rightarrow \text{ Não;} \\ 2 \rightarrow \text{ Sim;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

4. **Características socioeconômicas:**

- L01: indica qual a cor ou raça do indivíduo.

$$\text{L01} = \begin{cases} 1 \rightarrow \text{Branca;} \\ 2 \rightarrow \text{Preta;} \\ 3 \rightarrow \text{Amarela;} \\ 4 \rightarrow \text{Parda;} \\ 5 \rightarrow \text{Indígena;} \\ 6 \rightarrow \text{Outra;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

- L03: indica qual a religião do indivíduo.

$$\text{L03} = \begin{cases} 1 \rightarrow \text{Nenhuma;} \\ 2 \rightarrow \text{Evangélica/Protestante;} \\ 3 \rightarrow \text{Católica;} \\ 4 \rightarrow \text{Espírita;} \\ 5 \rightarrow \text{Judaísmo;} \\ 6 \rightarrow \text{Budismo;} \\ 7 \rightarrow \text{Umbanda/Candomblé;} \\ 8 \rightarrow \text{Islamismo;} \\ 9 \rightarrow \text{Outras;} \\ 99 \rightarrow \text{NS/NR.} \end{cases}$$

- L11: indica se o indivíduo tem filhos.

$$\text{L11} = \begin{cases} 1 \rightarrow \text{ Não;} \\ 2 \rightarrow \text{ Sim;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

- L13: indica se o indivíduo frequenta atualmente algum curso regular em escola ou universidade/faculdade.

$$\text{L13} = \begin{cases} 1 \rightarrow \text{ Não;} \\ 2 \rightarrow \text{ Sim;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

5. **Informações sobre a presença de animais:**

- P01: indica se o indivíduo possui algum animal em seu domicílio.

$$\text{P01} = \begin{cases} 1 \rightarrow \text{ Não;} \\ 2 \rightarrow \text{ Sim;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

6. **Variáveis criadas a partir do questionário:**

- IMC_cat: apresenta a relação entre o peso e a altura de um indivíduo. Seu cálculo é feito de acordo com a faixa etária do indivíduo e baseado no resultado obtido, o indivíduo é alocado em uma categoria.

$$\text{IMC\_cat} = \begin{cases} 0 \rightarrow \text{Baixo peso;} \\ 1 \rightarrow \text{Eutrofia;} \\ 2 \rightarrow \text{Sobrepeso;} \\ 3 \rightarrow \text{Obesidade.} \end{cases}$$

- Fumo2: indica em qual categoria de tabagismo o indivíduo se enquadra.

$$\text{Fumo2} = \begin{cases} 0 \rightarrow \text{Nunca fumou;} \\ 1 \rightarrow \text{Ex-fumante e fumante;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

- Alcool: indica em qual classificação do consumo de bebida alcoólica o indivíduo se enquadra.

$$\text{Alcool} = \begin{cases} 0 \rightarrow \text{Nunca bebeu;} \\ 1 \rightarrow \text{Parou de beber;} \\ 2 \rightarrow \text{Bebe atualmente;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

- Escolaridade_ind3: indica qual a categoria de escolaridade do indivíduo.

$$\text{Escolaridade\_ind3} = \begin{cases} 0 \rightarrow \leq 9 \text{ anos;} \\ 1 \rightarrow 10 \text{ a } 12 \text{ anos;} \\ 2 \rightarrow > 12 \text{ anos;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

- Conjugal3: indica qual a categoria da situação conjugal do indivíduo.

$$\text{Conjugal3} = \begin{cases} 1 \rightarrow \text{Casado/União Estável;} \\ 2 \rightarrow \text{Separado/Desquitado/Divorciado;} \\ 3 \rightarrow \text{Solteiro;} \\ 4 \rightarrow \text{Viúvo;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

- Probs15: indica se o indivíduo teve problema de saúde nos últimos 15 dias.

$$\text{Probs15} = \begin{cases} 1 \rightarrow \text{ Não;} \\ 2 \rightarrow \text{ Sim;} \\ 9 \rightarrow \text{NS/NR.} \end{cases}$$

7. **Banco de Atividade Física:**

- AF_GLOBAL_cat_OMS_1: indica se o indivíduo, de acordo com sua faixa etária, cumpre a recomendação de atividade física (em minutos por semana).

$$\text{AF\_GLOBAL\_cat\_OMS\_1} = \begin{cases} 0 \rightarrow \text{Não cumpre a recomendação;} \\ 1 \rightarrow \text{Cumpre a recomendação.} \end{cases}$$

Capítulo 3

Resultados

Neste capítulo apresentamos, primeiramente, uma breve análise dos dados, a fim de verificar o comportamento de cada covariável em relação as categorias da variável resposta, sendo eles, ausência de TMC e presença de TMC. Em seguida, temos a aplicação dos métodos de classificação binária, sendo eles, regressão logística, árvores de classificação e florestas aleatórias utilizados para estudar a prevalência de transtornos mentais comuns (TMC) e os diferentes pontos de cortes obtidos para cada um deles. A fim de comparar os resultados dos métodos aplicados adotamos algumas medidas de desempenho.

3.1 Análise Descritiva e Exploratória dos Dados

Para verificar o comportamento das variáveis do banco de dados estudado, realizamos algumas análises descritivas a fim de entender o comportamento das covariáveis selecionadas nos diferentes níveis da variável resposta. As análises foram feitas utilizando o software `R` (R Development Core Team, 2023) (ver Apêndice A).

Na Tabela 3.1, averiguamos a quantidade de observações faltantes em cada covariável do banco de dados. Vale ressaltar que algumas das covariáveis selecionadas poderiam assumir como resposta NS/NR (Não Sabe/Não Respondeu), como isso não nos traz uma informação relevante optamos por tratar essas repostas como uma informação faltante.

Notamos que existem dados faltantes em algumas covariáveis, porém como nenhuma das covariáveis possui uma grande quantidade de informações faltantes comparada ao número total de observações (3722 indivíduos) optamos por manter todas as covariáveis e retirar as observações faltantes em cada uma delas. Sendo assim, as análises gráficas e ajuste dos modelos foram realizados desconsiderando as observações faltantes, totalizando

3373 observações.

Tabela 3.1: Tabela da quantidade de informações faltantes em cada variável.

Variável	**Quantidade de Informações Faltantes**
idade	0
sexo	0
F101	3
k201a	3
k205a_hrs	17
k205b_hrs	20
k206_hrs	11
k207_hrs	13
k208_hrs	11
k209_hrs	12
k210	2
L01	20
L03	15
L11	14
L13	10
p01	139
imc_cat	90
fumo2	3
alcool	14
escolaridade_ind3	16
AF_GLOBAL_cat_OMS_1	37
conjugal3	9
probs15	4

Inicialmente, na Figura 3.1, observamos o comportamento da variável resposta através de um gráfico de barras, no qual é possível notar que a maioria dos indivíduos, em torno de 77%, não apresenta TMC.

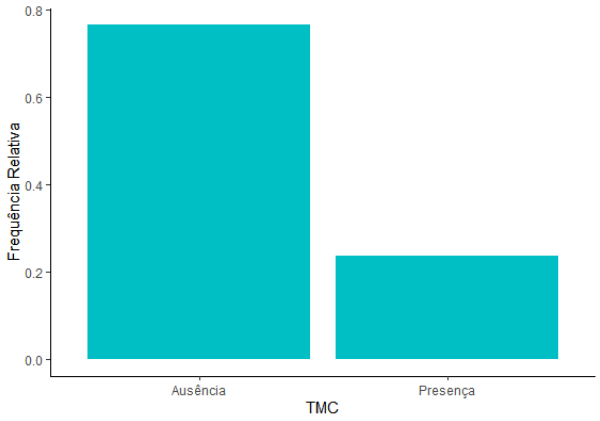
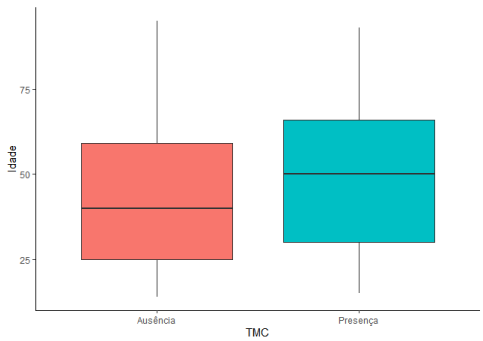


Figura 3.1: Gráfico de barras da frequência relativa da variável resposta.

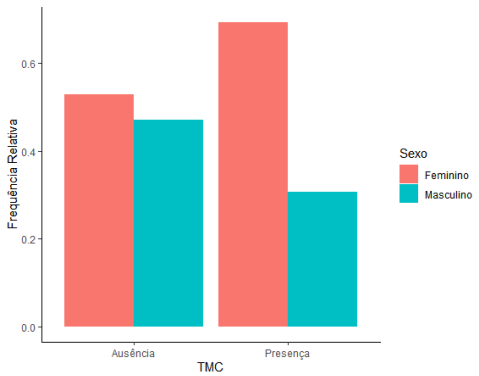
Em seguida, realizamos a análise gráfica para cada uma das covariáveis em relação às diferentes categorias da variável resposta (Ausência de TMC e Presença de TMC) e apresentamos essa análise respeitando os blocos em que cada covária vel foi alocada no banco de dados como descrito anteriormente (Seção 2.6).

Para as variáveis de identificação (Figura 3.2), podemos observar através da Figura 3.2a que a mediana da variável idade é maior para os indivíduos com presença de TMC. No entanto, o limites superior e inferior, assim como a distância interquartílica, são bem próximos para as duas classes da variável resposta. Na Figura 3.2b, notamos que, para a ausência de TMC, a frequência relativa dos indivíduos do sexo Feminino é bem próxima do sexo Masculino. Porém, esse padrão muda quando analisamos indivíduos com presença de TMC, em que a frequência relativa do sexo Feminino é muito superior à do sexo Masculino, dando um indicativo que a variável sexo terá uma forte influência na variável resposta.

1. **Variáveis de Identificação:**



(a) Boxplot da variável Idade.



(b) Gráfico de barras da variável Sexo.

Figura 3.2: Gráficos referentes ao bloco Variáveis de Identificação.

No que diz respeito à qualidade de vida, observamos a variável F101, que traz informações sobre o estado de saúde declarado pelo indivíduo. Pela Figura 3.3, notamos que a maioria dos indivíduos que não possuem TMC declara seu estado de saúde como “Boa”. No entanto, para aqueles com presença de TMC, a maior parte declara seu estado de saúde como “Regular”. Além disso, podemos destacar que indivíduos que possuem transtorno mental comum declaram com maior frequência o estado de saúde como “Ruim” ou “Muito Ruim”, quando comparados aos indivíduos sem transtorno mental comum. O oposto acontece para o estado “Excelente” ou “Muito Bom”, ou seja, os indivíduos com ausência de TMC declaram com maior frequência esse estado do que aqueles com presença de TMC. A partir disso, temos um indicativo de que os indivíduos com TMC aparentam ter consciência do seu estado atual de saúde, já que, em sua maioria, o declaram como“Regular” ou “Ruim”. No entanto, notamos que cerca de 39% dos indivíduos com TMC declaram seu estado de saúde como “Excelente/Muito boa” ou “Boa”. Sendo assim, esses indivíduos aparentam não ter plena consciência do seu estado atual de saúde.

2. **Qualidade de vida:**

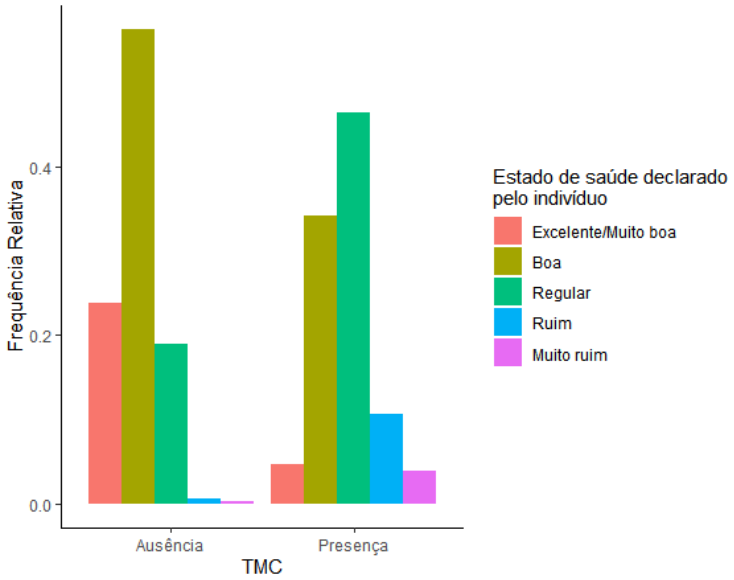


Figura 3.3: Gráfico de barras da variável que indica como o indivíduo diria que está seu estado de saúde.

No Banco de Atividade Física analisamos a Figura 3.4, que diz respeito a variável criada para averiguar se o indivíduo está seguindo a recomendação, de acordo com sua faixa etária, de minutos semanais da prática de atividade física. Observamos que para os dois níveis da variável resposta a maioria dos indivíduos cumpre a recomendação. Vale

ressaltar que a frequência relativa para os dois níveis é muito maior para os indivíduos que cumprem a recomendação do que para aqueles que não cumprem.

3. **Bloco de Atividade Física:**

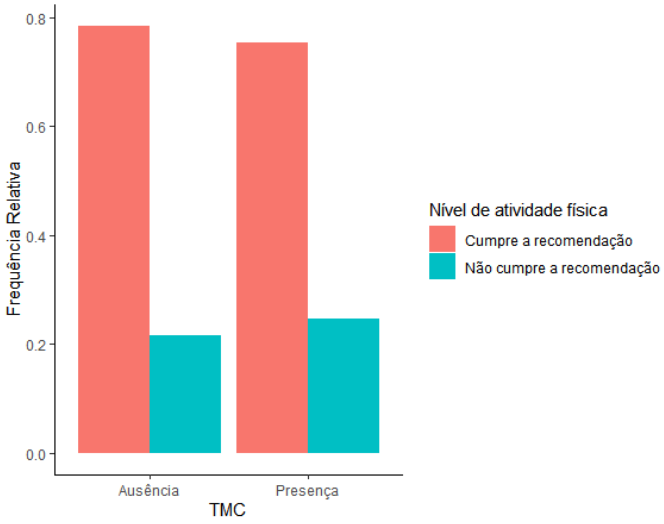


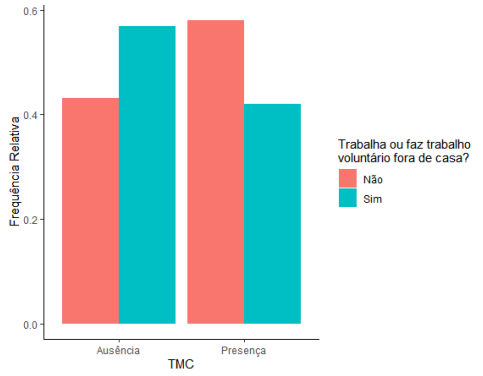
Figura 3.4: Gráfico de barras da variável que indica se o indivíduo cumpre a recomendação de atividade física (em minutos por semana).

Agora, no bloco de Atividade física (Figura 3.5), observamos na Figura 3.5a que a maioria dos indivíduos que não apresentam TMC trabalham ou fazem trabalho voluntário fora de casa, já para aqueles com TMC a maioria não trabalha ou faz trabalho voluntário fora de casa, dessa forma, temos um indicativo de que a variável K201a terá influência na variável resposta.

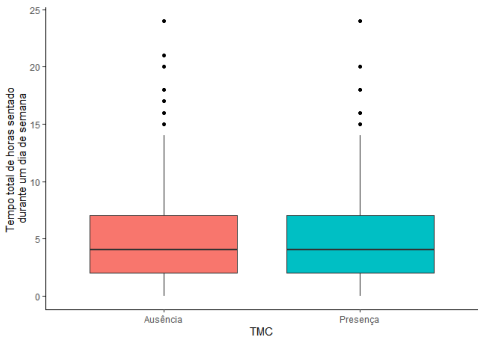
Em relação ao total de horas sentado durante um dia de semana e um dia de final de semana, Figuras 3.5b e 3.5c, respectivamente, não observamos diferenças significativas entre as diferentes classes da variável resposta, uma vez que, ambos apresentam outliers, medianas similares e intersecção entre as caixas. As Figuras 3.5f e 3.5g também são muito semelhantes entre si para indivíduos com e sem TMC. Dessa forma, temos um indicativo de que o total de horas sentado, assim como o total de horas no computador, durante um dia de semana e um dia de final de semana não terá influência sobre a variável resposta.

Por fim vemos, pela Figura 3.5h, que em ambas as classes da variável resposta a maioria dos indivíduos não pratica regularmente algum tipo de exercício físico ou esporte, vale ressaltar que a frequência de indivíduos que não praticam regularmente algum tipo de exercício físico é ainda maior para aqueles com presença de TMC.

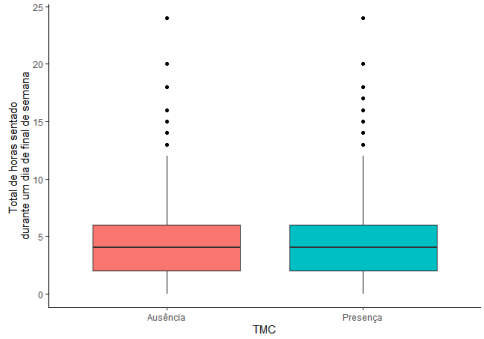
4. **Atividade física:**



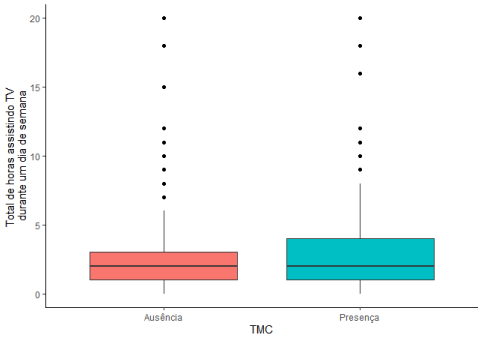
(a) Gráfico de barras da variável que indica se atualmente o indivíduo trabalha fora de casa.



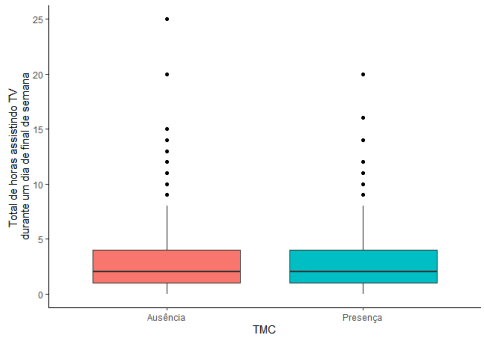
(b) Boxplot do tempo total (em horas) gasto sentado durante um dia de semana.



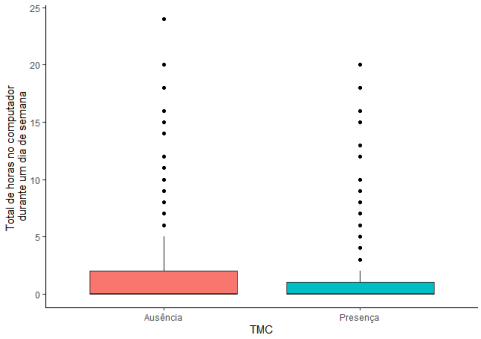
(c) Boxplot do tempo total (em horas) gasto sentado durante um dia de final de semana.



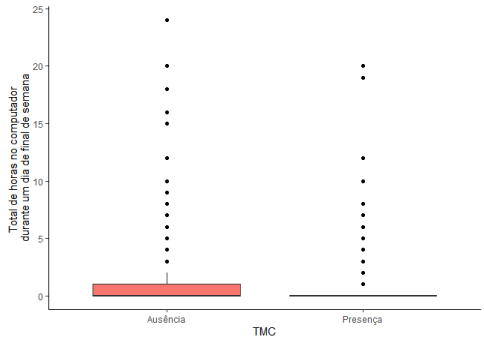
(d) Boxplot do tempo total (em horas) gasto assistindo TV durante um dia de semana.



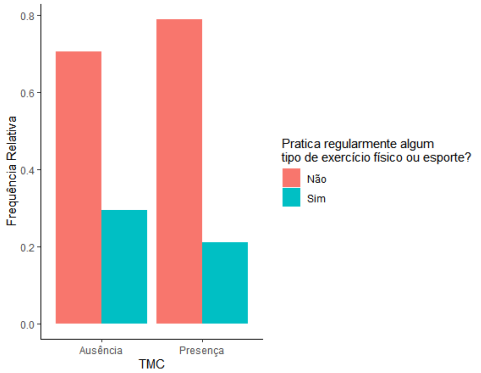
(e) Boxplot do tempo total (em horas) gasto assistindo TV durante um dia de final de semana.



(f) Boxplot do tempo total (em horas) gato no computador durante um dia de semana.



(g) Boxplot do tempo total (em horas) gasto no computador durante um dia de final de semana.

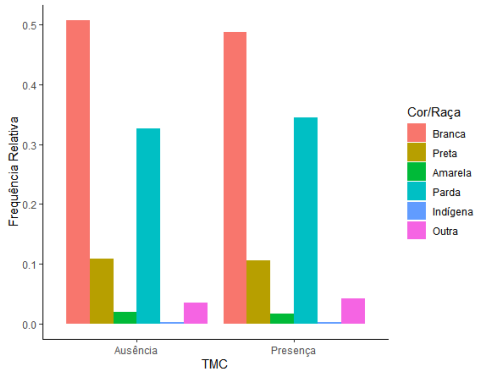


(h) Gráfico de barras da variável que indica a pratica regular de algum tipo de exercício físico.

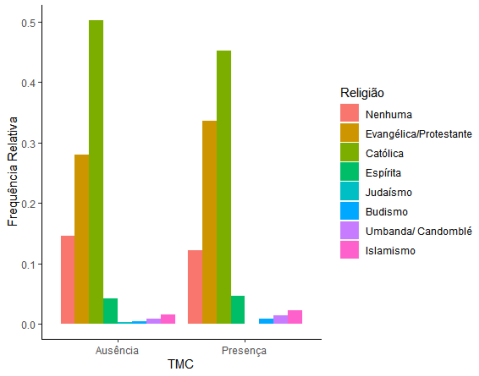
Figura 3.5: Gráficos referente ao bloco Atividade física.

Observando o bloco que contempla as características socioeconômicas (Figura 3.6) notamos um comportamento similar para todas as variáveis quando observamos os indivíduos com TMC ausênte e presente. Sendo assim, temos um indicativo que as variáveis relacionadas as características socioeconômicas não terão influência significativa na classificação da variável resposta.

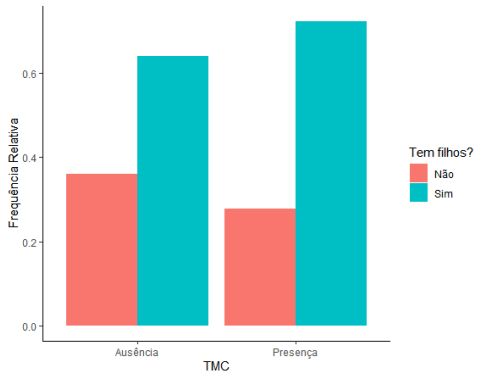
5. **Características socioeconômicas:**



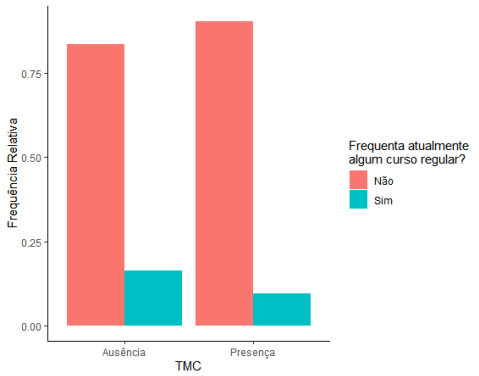
(a) Gráfico de barras da variável que indica a cor ou raça do indivíduo.



(b) Gráfico de barras da variável que indica a religião do indivíduo.



(c) Gráfico de barras da variável que indica se o indivíduo possui filhos.



(d) Gráfico de barras da variável que indica se o indivíduo frequenta algum curso regular.

Figura 3.6: Gráficos referente ao bloco Características socioeconômicas.

Analisando a presença de animais no domicílio (Figura 3.7) notamos que a maioria dos indivíduos, com e sem TMC, não possuem animais em seu domicílio, porém quando temos a presença de TMC a frequência relativa dos indivíduos que não possuem animal em seu domicílio é mais próxima daqueles que possuem.

6. **Informações sobre a presença de animais:**

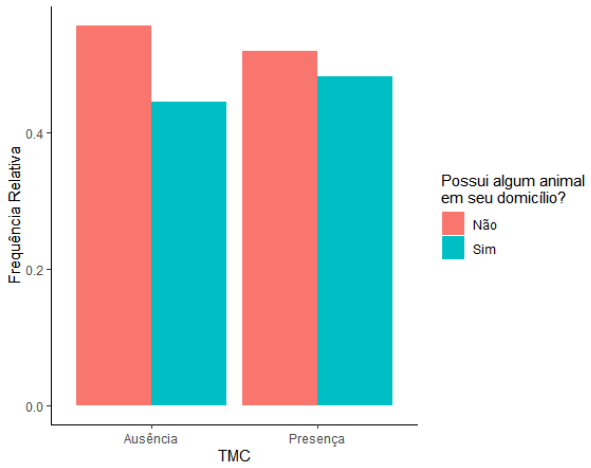


Figura 3.7: Gráfico de barras da variável que indica se o indivíduo possui algum animal em seu domicílio.

Pelas variáveis criadas a partir do questionário (Figura 3.8) notamos na Figura 3.8a que, quando comparamos os indivíduos com TMC em relação aos indivíduos sem TMC, há um aumento da frequência relativa dos indivíduos com baixo peso ou com obesidade enquanto a frequência de indivíduos com eutrofia ou sobrepeso diminui, mesmo que ainda sejam valores próximos para as duas categorias da variável resposta. Para a variável Fumo2 (Figura 3.8b) observamos que a categoria ex-fumante e fumante tem maior frequência para indivíduos com presença de TMC quando comparados aqueles com ausência e o oposto ocorre para a categoria nunca fumou.

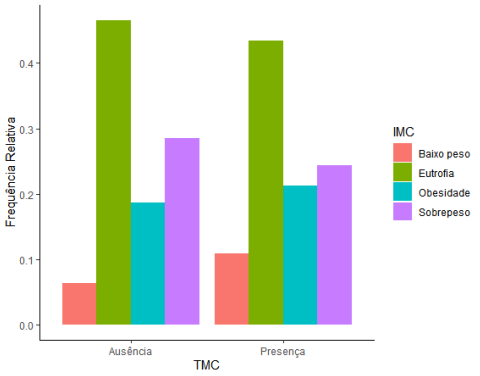
Na Figura 3.8c podemos notar que, em relação aos indivíduos sem transtornos mentais comuns, os indivíduos com TMC apresentam uma diminuição na categoria bebe atualmente e um aumento para parou de beber, isso pode estar relacionado à alguns transtornos mentais comuns serem tratados com medicações que não devem ter interação com bebidas alcoólicas.

Observando o nível de escolaridade (Figura 3.8d) vemos que indivíduos com menor escolaridade (≤ 9 anos) são a maioria quando analisado a presença de TMC enquanto que indivíduos com escolaridade de 10 a 12 anos são a maior parte para ausência de TMC, sendo assim, temos um indicativo de que a variável Escolaridade_ind3 pode apresentar uma influência significativa na variável resposta. Já na Figura 3.8e, para os dois níveis da variável resposta a maior parte dos indivíduos tem situação conjugal como Casado/União Estável, porém vemos um aumento na frequência de indivíduos Separado/Desquitado/Divorciado e Viúvo quando analisamos aqueles com presença de TMC.

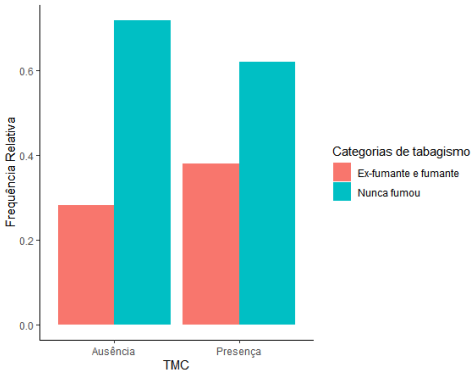
Por fim, para a variável Probs15, apresentada na Figura 3.8f, os indivíduos sem TMC,

em sua maioria, não apresentaram problema de saúde nos últimos 15 dias, o mesmo ocorre para aqueles com TMC mas, nesse caso, os indivíduos que apresentaram problema de saúde nos últimos 15 dias são mais frequentes quando comparados a outra classe.

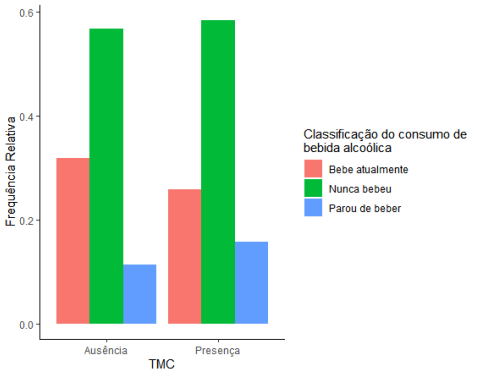
7. **Variáveis criadas a partir do questionário:**



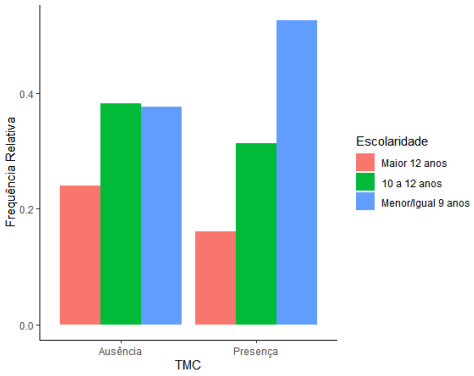
(a) Gráfico de barras da variável IMC categorizada.



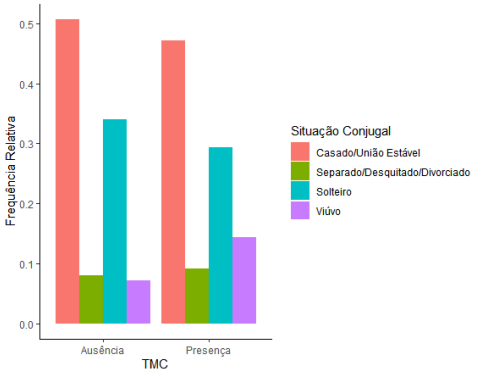
(b) Gráfico de barras da variável que indica a categoria de tabagismo.



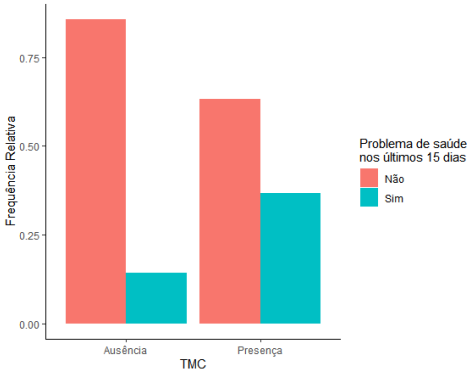
(c) Gráfico de barras da variável que indica a classificação do consumo de bebida alcoólica.



(d) Gráfico de barras da variável que indica a categoria de escolaridade.



(e) Gráfico de barras da variável que indica a categoria da situação conjugal.



(f) Gráfico de barras da variável que indica problemas de saúde nos últimos 15 dias.

Figura 3.8: Gráficos referente ao bloco Variáveis criadas a partir do questionário.

3.2 Divisão do Banco de Dados

Inicialmente, realizamos o *data-splitting* (ver Apêndice B), ou seja, dividimos de forma aleatória nosso conjunto de dados em dois subconjuntos distintos, sendo eles, conjunto de treinamento (70% dos dados) e conjunto de validação (30% dos dados). A partir da Tabela 3.2 notamos que a proporção de indivíduos com ausência e presença de transtornos mentais comuns se mantém constante para o conjunto de treinamento e validação.

Dessa forma, utilizaremos o conjunto de treinamento para estimar as funções de predição em cada um dos métodos de classificação binária, e o conjunto de validação será usado para avaliar o poder preditivo dos mesmos.

Tabela 3.2: Proporção das categorias da variável resposta nos conjuntos de treinamento e validação.

TMC	**Proporção**	
	Conjunto de Treinamento	Conjunto de Validação
Ausência	0,77	0,77
Presença	0,23	0,23

3.3 Árvores de Classificação

Nesta seção, apresentamos os resultados obtidos por meio da implementação das árvores de classificação utilizando o `R` (R Development Core Team, 2023) (ver Apêndice C). Sendo assim, ajustamos a árvore de classificação e realizamos sua poda, como detalhado na Subseção 2.1.1, obtendo a árvore de classificação exibida na Figura 3.9.

Pela Figura 3.9 notamos que três variáveis foram escolhidas pela árvore de classificação, sendo elas:

- F101: indica como o indivíduo classifica seu estado de saúde;
- probs15: indica se o indivíduo teve problemas de saúde nos últimos 15 dias;
- k205b_hrs: indica o tempo total (em horas) que o indivíduo passa sentado durante um dia de final de semana.

Interpretando a Figura 3.9 observamos que a maioria dos indivíduos não possui TMC, pois os agrupamentos em que existem indivíduos com presença de TMC chega ao máximo de 8% de representatividade (por classe) para toda a base. Dessa forma, o agrupamento

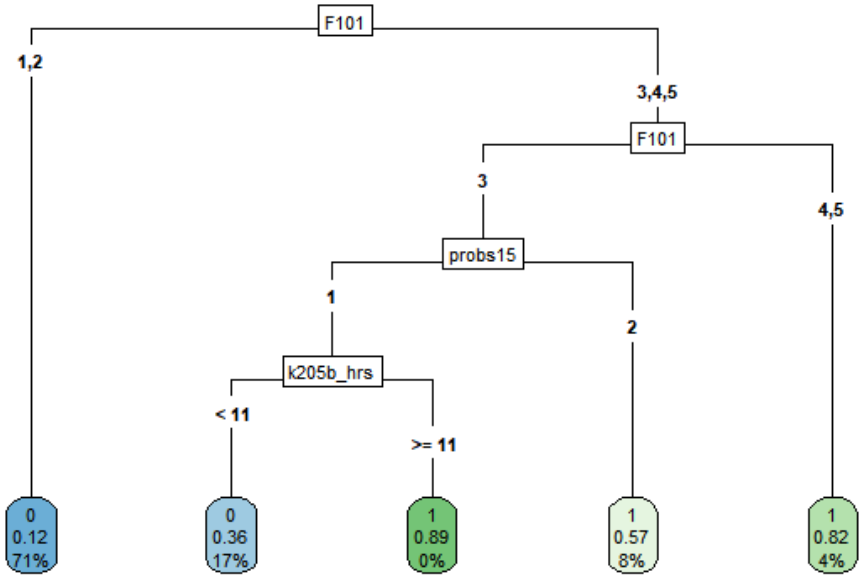


Figura 3.9: Árvore de Classificação.

com maior presença de indivíduos com TMC é obtido quando o indivíduo classifica seu estado de saúde como regular e teve problemas de saúde nos últimos 15 dias.

Analisando cada uma das folhas temos que para a primeira folha, apenas é utilizada a variável que indica como o indivíduo classifica seu estado de saúde (F101), sendo assim, 71% dos indivíduos do conjunto de treinamento classificam seu estado de saúde como excelente/muito boa ou boa com 12% de probabilidade de possuírem TMC, desse modo, indivíduos que estão nesse nó são classificados com ausência de TMC. O segundo nó engloba indivíduos que classificam seu estado de saúde como regular, não tiveram problemas de saúde nos últimos 15 dias e passam menos de 11 horas sentado durante um dia de final de semana, sendo assim 17% dos indivíduos do conjunto de treinamento se enquadram nessa condição com 36% de probabilidade de possuírem TMC, desse modo, indivíduos que estão nesse nó são classificados com ausência de TMC. Por outro lado, no terceiro nó temos indivíduos que classificam seu estado de saúde como regular, não tiveram problemas de saúde nos últimos 15 dias e passam 11 ou mais horas sentado durante um dia de final de semana, 0% dos indivíduos do conjunto de treinamento se enquadram nessa condição com 89% de probabilidade de possuírem TMC, então esses indivíduos são classificados com presença de TMC. Ademais, a quarta folha é composta por indivíduos que classificam seu estado de saúde como regular e tiveram problemas de saúde nos últimos 15 dias, sendo assim 8% dos indivíduos do conjunto de treinamento se enquadram nessa

condição com 57% de probabilidade de possuírem TMC, indivíduos que estão nesse nó são classificados com presença de TMC. Por fim, a quinta folha é composta por indivíduos que classificam seu estado de saúde como ruim ou muito ruim, 4% dos indivíduos do conjunto de treinamento se enquadram nessa condição com 82% de probabilidade de possuírem TMC, sendo assim, indivíduos que estão nesse nó são classificados com presença de TMC.

A partir do ajuste da árvore de classificação construímos uma matriz de confusão (Tabela 3.3) na qual podemos comparar as previsões obtidas com os valores reais. Primeiramente, observamos as previsões obtidas para um ponto de corte = 0,5, ou seja, uma observação será atribuída à classe positiva se a probabilidade prevista for maior que 0,5 e à classe negativa se for igual ou menor que 0,5. Sendo assim, o indivíduo será classificado com presença de TMC se a probabilidade prevista for maior que 0,5 e com ausência de TMC se for igual ou menor que 0,5.

Em seguida, a fim de otimizar o desempenho da árvore de classificação, utilizamos a análise da curva ROC, apresentada na Figuras 3.10, e calculamos os pontos de corte descritos na Seção 2.5, dessa forma, obtivemos Média-G = $0,24$ e J = $0,24$. Sendo assim, construímos a matriz de confusão para os diferentes pontos de corte (Tabela 3.3) e apresentamos as medidas de desempenho para os mesmos (Tabela 3.4).

Tabela 3.3: Matriz de Confusão para Árvore de Classificação.

Valor Predito	Valor Verdadeiro			
	Ponto de corte = 0,5		Ponto de corte = 0,24	
	Y = 0	Y = 1	Y = 0	Y = 1
Y = 0	0,74	0,15	0,61	0,09
Y = 1	0,03	0,08	0,16	0,14

Pela Tabela 3.3, notamos que, para o ponto de corte = 0,5 o modelo classificou corretamente 8% dos casos como positivos (presença de TMC) e 74% dos casos como negativos (ausência de TMC). Houve 3% falsos positivos, ou seja, casos previstos como positivos mas que eram negativos na realidade e 15% falsos negativos, indicando casos que foram erroneamente classificados como negativos. Já para o ponto de corte = 0,24 o modelo classificou corretamente 14% dos casos como positivos e 61% dos casos com negativos, no entanto, houve um aumento de falsos positivos e uma redução de falsos negativos.

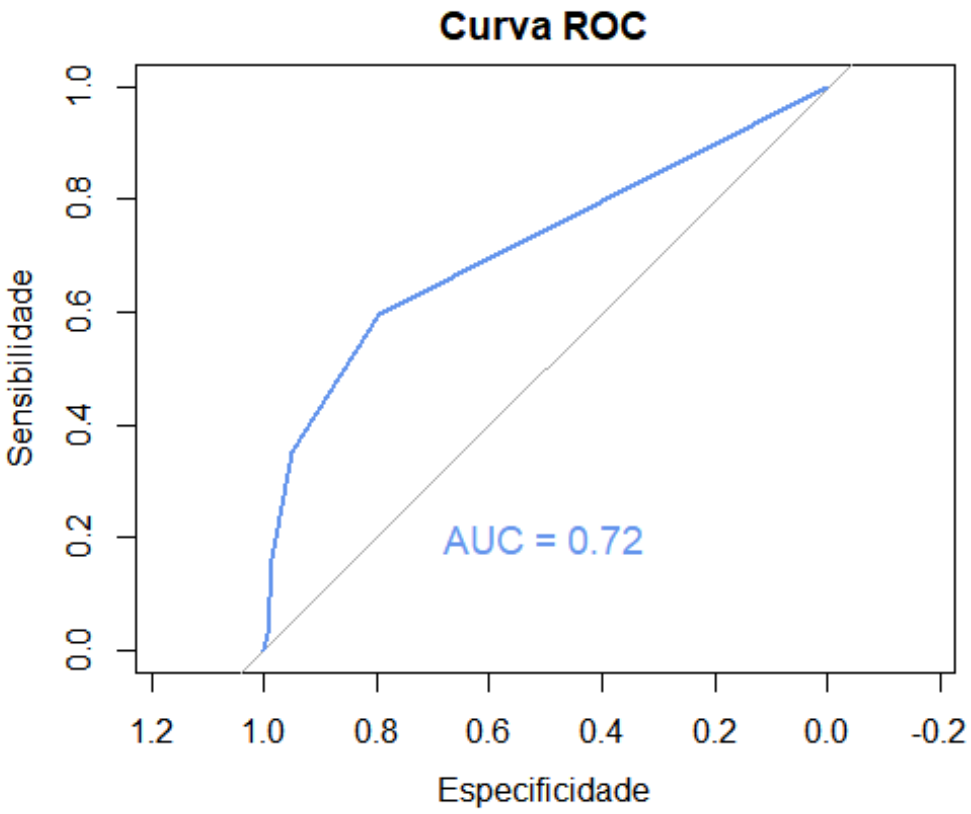


Figura 3.10: Curva ROC

A Tabela 3.4 apresenta as medidas de desempenho para a árvore de classificação ajustada. Para o ponto de corte padrão de 0,5, a acurácia é de 81%, indicando a proporção total de previsões corretas. No entanto, ao ajustar o ponto de corte para 0.24, a acurácia diminui para 74%. Quanto à sensibilidade, que representa a capacidade do modelo em identificar positivos verdadeiros, observa-se um aumento significativo de 32% (ponto de corte 0,5) para 59% (ponto de corte 0,24). Em contrapartida, a especificidade, que mede a capacidade do modelo em identificar negativos verdadeiros, diminui de 95% (ponto de corte 0,5) para 79% (ponto de corte 0,24).

A precisão, indicando a proporção de positivos previstos corretamente, é de 69% para o ponto de corte de 0,5, mas diminui para 46% com o ponto de corte ajustado para 0,24. O valor predito negativo, que representa a proporção de negativos verdadeiros entre as previsões negativas, aumenta de 82% para 86%. A estatística F1, que combina precisão e sensibilidade, mostra uma melhoria de 44% para 52%.

Tabela 3.4: Medidas de Desempenho para Árvore de Classificação.

Medidas de Desempenho		
Ponto de Corte	**0,5**	**0,24**
Acurácia	0,81	0,74
Sensibilidade	0,32	0,59
Especificidade	0,95	0,79
Precisão	0,69	0,46
Valor Predito Negativo	0,82	0,86
Estatística F1	0,44	0,52

3.4 Florestas Aleatórias

Agora, apresentamos os resultados obtidos por meio do ajuste de florestas aleatórias utilizando o `R` (R Development Core Team, 2023) (ver Apêndice D). O ajuste da floresta aleatória foi realizado como detalhado na Subsecção 2.2.

Pela Figura 3.11, podemos observar a importância de cada uma das variáveis no ajuste da floresta aleatória. Dessa forma, destacamos que as seis variáveis com maior importância são:

- F101: indica como o indivíduo classifica seu estado de saúde;
- idade: idade do indivíduo;
- k205a_hrs: indica o tempo total (em horas) que o indivíduo gasta sentado durante um dia de semana;
- k205b_hrs: indica o tempo total (em horas) que o indivíduo gasta sentado durante um dia de de final de semana;
- k207_hrs: indica o tempo total (em horas) que o indivíduo gasta assistindo TV durante um dia de final de semana.
- k206_hrs: indica o tempo total (em horas) que o indivíduo gasta assistindo TV durante um dia de semana.

Vale ressaltar, que duas das seis variáveis classificadas com maior importância pelo ajuste da floresta aleatória foram também selecionadas pelo ajuste da árvore de classificação, sendo elas, a variável que indica como o indivíduo classifica seu estado de saúde e o tempo total (em horas) que o indivíduo gasta sentado durante um dia de final de semana.

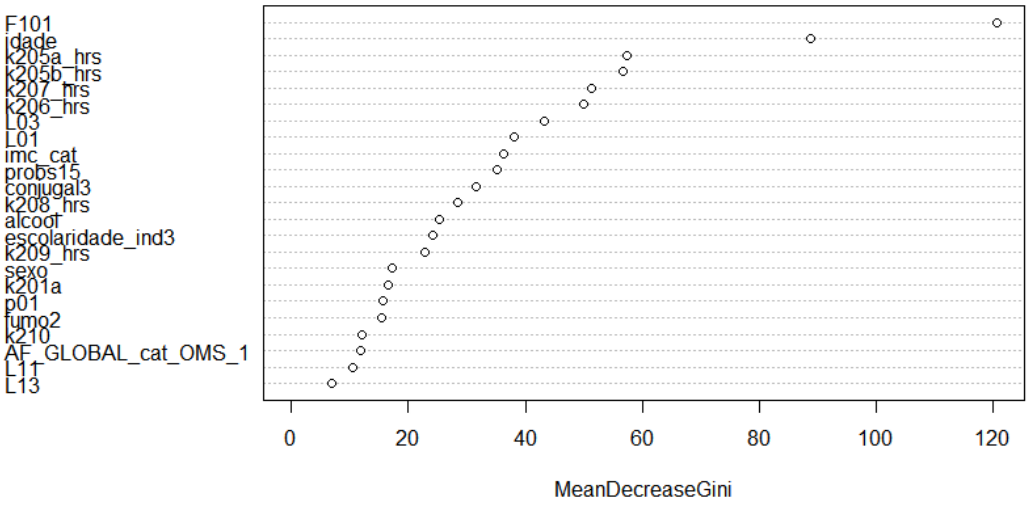


Figura 3.11: Gráfico de Importância das Variáveis para o ajuste da Floresta Aleatória.

A partir do ajuste da floresta aleatória calculamos os pontos de corte obtidos através das duas abordagens sugeridas: média geométrica e estatística de Youden a partir da análise da curva ROC (Figura 3.12). Dessa forma, apresentamos nas Tabelas 3.5 e 3.6, a matriz de confusão e as medidas de desempenho para os pontos de corte, sendo eles, Média-G $= 0,23$ e J $= 0,23$.

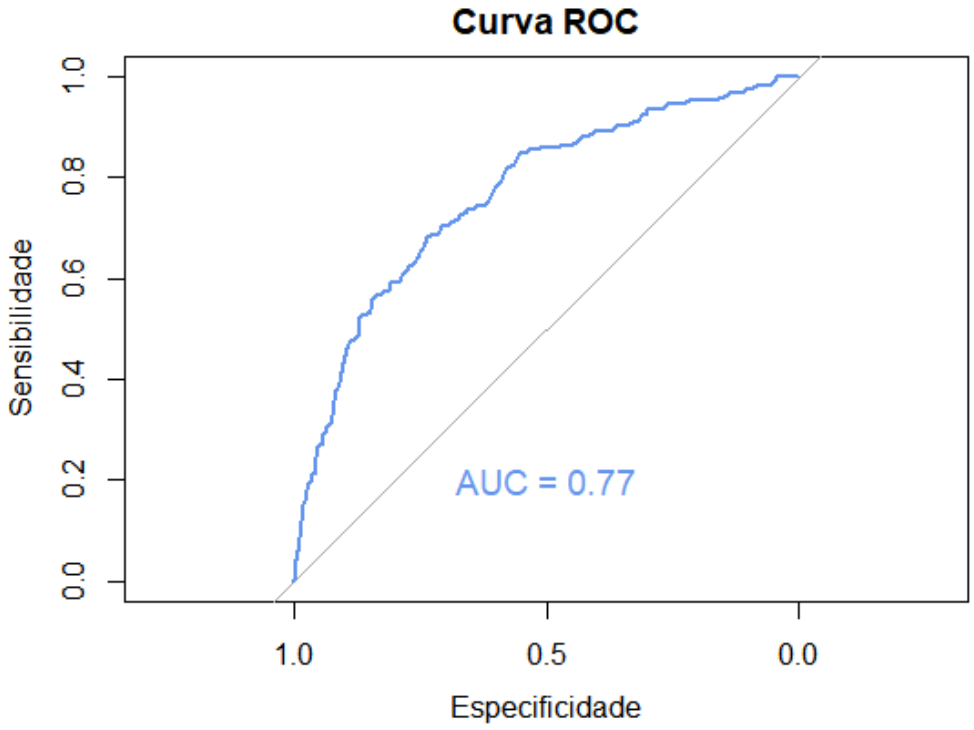


Figura 3.12: Curva ROC

Tabela 3.5: Matriz de Confusão para Floresta Aleatória.

Valor Predito	Valor Verdadeiro			
	Ponto de corte = 0,5		Ponto de corte = 0,23	
	Y = 0	Y = 1	Y = 0	Y = 1
Y = 0	0,74	0,17	0,57	0,07
Y = 1	0,03	0,06	0,20	0,16

Observando a Tabela 3.5, temos que, para o ponto de corte = 0,5 o modelo classificou corretamente 6% dos casos como positivos (presença de TMC) e 74% dos casos como negativos (ausência de TMC). Houve 3% falsos positivos, ou seja, casos previstos como positivos mas que eram negativos na realidade e 17% falsos negativos, indicando casos que foram erroneamente classificados como negativos. Já para o ponto de corte = 0,23 o modelo classificou corretamente 16% dos casos como positivos e 57% dos casos com negativos, no entanto, houve um aumento de falsos positivos e uma redução de falsos negativos.

Tabela 3.6: Medidas de Desempenho para Floresta Aleatória.

Medidas de Desempenho		
Ponto de Corte	**0,5**	**0,23**
Acurácia	0,79	0,72
Sensibilidade	0,25	0,68
Especificidade	0,95	0,73
Precisão	0,63	0,43
Valor Predito Negativo	0,81	0,88
Estatística F1	0,36	0,53

Pela Tabela 3.6, analisamos que para o ponto de corte padrão de 0,5, a acurácia é de 79%, indicando a proporção total de previsões corretas. A sensibilidade, que representa a capacidade do modelo em identificar positivos verdadeiros, é de 25%, sugerindo uma limitação na capacidade de detectar casos positivos. A especificidade, medindo a habilidade do modelo em identificar verdadeiros negativos, é alta, alcançando 95%. A precisão, indicando a proporção de positivos previstos corretamente, é de 63%.

Ao ajustar o ponto de corte para 0,23, a acurácia diminui para 72%. No entanto, a sensibilidade melhora significativamente para 68%, indicando uma capacidade aprimorada em identificar casos positivos. A especificidade, embora reduza para 73%, ainda é relativamente alta. A precisão diminui para 43%, sugerindo um compromisso entre sensibilidade e precisão.

O valor predito negativo aumenta de 81% para 88%, indicando uma maior capacidade de identificar corretamente casos negativos. A estatística F1, que combina precisão e sensibilidade, melhora de 36% para 53% com o ponto de corte ajustado.

Dessa forma, notamos que os resultados obtidos no ajuste da Floresta Aleatória são muito próximos daqueles obtidos no ajuste da Árvore de Classificação, e nos dois cenários o ponto de corte proposto utilizando as métricas Média-G e J apresentaram melhores resultados, uma vez que, promove um melhor equilíbrio entre sensibilidade e especificidade.

3.5 Regressão logística

Conforme descrito na Seção 2.3, implementamos a Regressão Logística utilizando todas as covariáveis presentes no conjunto de dados a fim de prever a variável resposta Y, que assume 1 para indivíduos com presença de TMC e 0 caso contrário. Com esse propósito, planejamos realizar a Regressão Logística considerando todas as variáveis independentes disponíveis no conjunto de dados. Adicionalmente, realizamos o mesmo procedimento utilizando apenas as covariáveis selecionadas pela técnica *Stepwise*. Ambos os ajustes serão executados utilizando a linguagem de programação R (R Development Core Team, 2023) (ver Apêndice E).

Primeiramente, considerando todas as covariáveis presentes no conjunto de dados, obtivemos, a partir do ajuste de uma regressão logística, as previsões obtidas para um ponto de corte = 0,5. Posteriormente, empregamos a análise da curva ROC, ilustrada na Figura 3.13. Nesse contexto, determinamos os pontos de corte conforme descrito na Seção

2.5, resultando em valores de Média-G = 0, 22 e J = 0, 22. Com base nesses resultados, elaboramos a matriz de confusão para diferentes pontos de corte (Tabela 3.7) e expomos as métricas de desempenho associadas a cada um deles (Tabela 3.10).

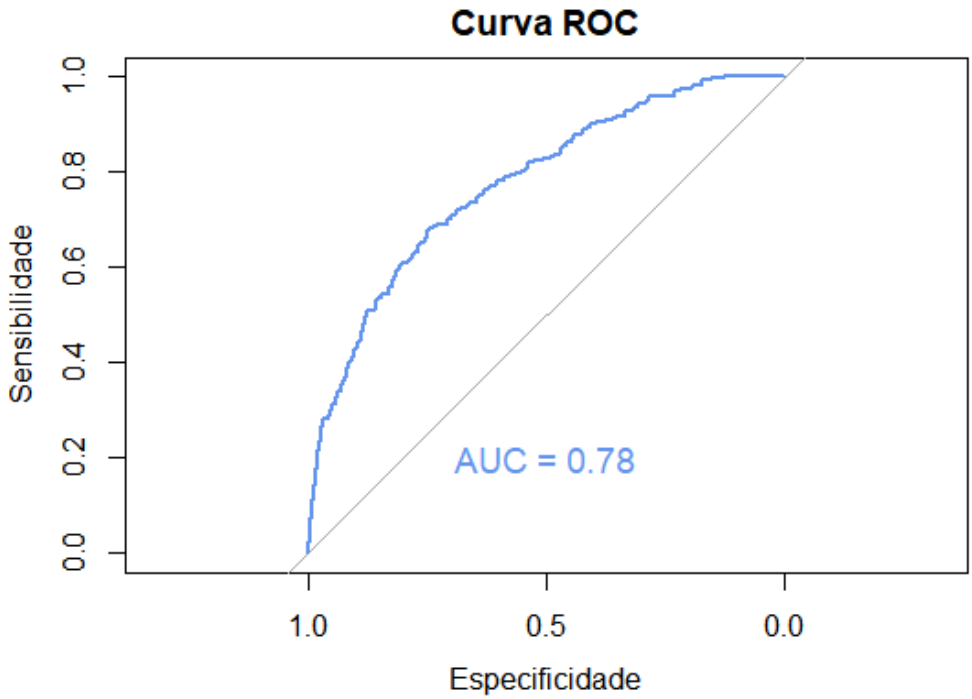


Figura 3.13: Curva ROC

Tabela 3.7: Matriz de Confusão para Regressão Logística.

Valor Predito	Valor Verdadeiro			
	Ponto de corte = 0,5		Ponto de corte = 0,22	
	Y = 0	Y = 1	Y = 0	Y = 1
Y = 0	0,72	0,16	0,58	0,07
Y = 1	0,05	0,07	0,19	0,16

Pela Tabela 3.7, temos que, para o ponto de corte = 0,5 o modelo classificou corretamente 7% dos casos como positivos (presença de TMC) e 72% casos como negativos (ausência de TMC). Houve 5% falsos positivos, ou seja, casos previstos como positivos mas que eram negativos na realidade e 16% falsos negativos, indicando casos que foram erroneamente classificados como negativos. Já para o ponto de corte = 0,22 o modelo classificou corretamente 16% casos como positivos e 58% casos com negativos, no entanto, houve um aumento de falsos positivos e uma redução de falsos negativos.

Tabela 3.8: Medidas de Desempenho para Regressão Logística.

Medidas de Desempenho		
Ponto de Corte	**0,5**	**0,22**
Acurácia	0,79	0,73
Sensibilidade	0,32	0,68
Especificidade	0,94	0,74
Precisão	0,61	0,44
Valor Predito Negativo	0,82	0,88
Estatística F1	0,42	0,53

Observamos pela Tabela 3.10 que para o ponto de corte padrão de 0,5, o modelo atinge uma acurácia de 79%, com uma sensibilidade de 32% e uma alta especificidade de 94%. A precisão, indicando a proporção de positivos previstos corretamente, é de 61%.

Ao ajustar o ponto de corte para 0,22, a acurácia diminui para 73%, já a sensibilidade melhora para 68%. A especificidade reduz para 74%, ainda mantendo um nível razoável. A precisão diminui para 44%, indicando um compromisso entre sensibilidade e precisão.

O valor predito negativo aumenta de 82% para 88%, indicando uma melhor capacidade do modelo em identificar corretamente casos negativos com o ponto de corte ajustado. A estatística F1, que combina precisão e sensibilidade, melhora de 42% para 53%.

Após realizado o ajuste da regressão logística com todas as covariáveis presente no conjunto de dados, implementamos a regressão logística utilizando as covariáveis selecionadas pelo método *Stepwise*. Desse modo, as covariáveis selecionadas foram:

- Sexo: indica o sexo do indivíduo;
- F101: indica como o indivíduo classifica seu estado de saúde;
- K201a: indica se atualmente o indivíduo trabalha ou faz trabalho voluntário fora de sua casa;

- k205b_hrs: indica o tempo total (em horas) que o indivíduo gasta sentado durante um dia de de final de semana;
- Fumo2: indica em qual categoria de tabagismo o indivíduo se enquadra;
- Conjugal3: indica qual a categoria da situação conjugal do indivíduo;
- probs15: indica se o indivíduo teve problemas de saúde nos últimos 15 dias.

Notamos assim, que das variáveis selecionadas pela árvore de classificação (F101, probs15 e K205b_hrs) todas também foram selecionadas pelo método *Stepwise* e das 6 variáveis classificadas com maior importancia pela floresta aleatória, duas foram selecionadas pelo método *Stepwise*, sendo elas, F101 e K205b_hrs.

Depois de ajustarmos o modelo apenas com as covariáveis selecionadas, efetuamos predições nas observações do conjunto de validação. Assim, ajustamos a Curva ROC, em que são baseadas as métricas Média-G e J, na Figura 3.14.

Dessa forma, obtivemos os pontos de corte Média-G $= 0,24$ e J $= 0,24$ e construímos para os mesmos suas respectivas matrizes de confusão (Tabela 3.9) e medidas de desempenho (Tabela 3.10).

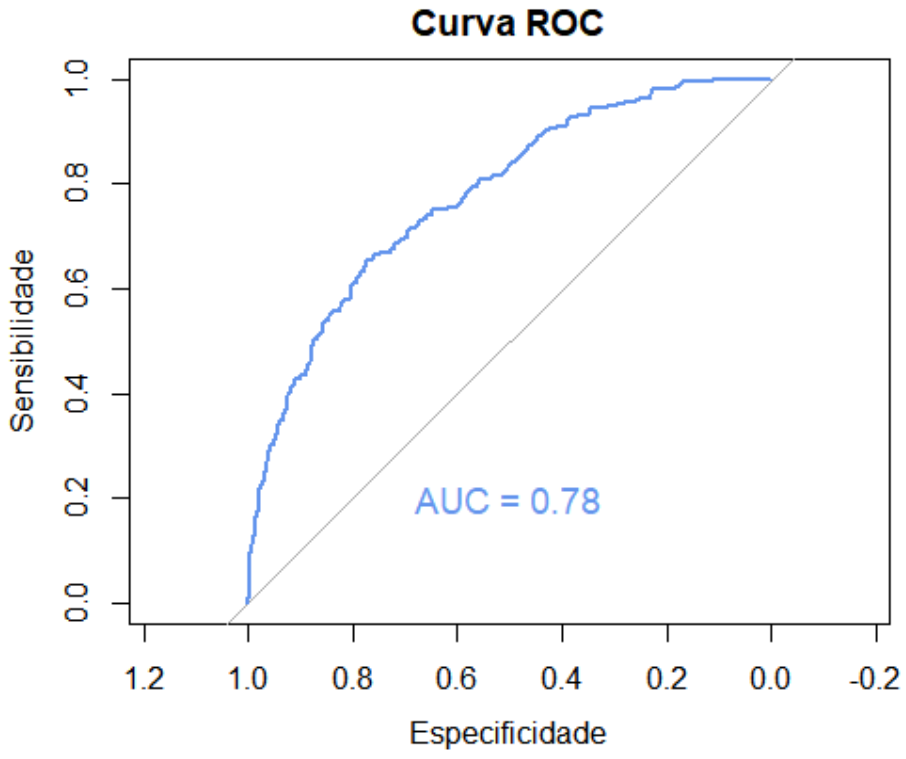


Figura 3.14: Curva ROC

Tabela 3.9: Matriz de Confusão para Regressão Logística.

Valor Predito	Valor Verdadeiro			
	Ponto de corte = 0,5		Ponto de corte = 0,24	
	Y = 0	Y = 1	Y = 0	Y = 1
Y = 0	0,73	0,16	0,60	0,08
Y = 1	0,04	0,07	0,17	0,15

Pela Tabela 3.9, temos que, para o ponto de corte = 0,5 o modelo classificou corretamente 7% dos casos como positivos (presença de TMC) e 0,73% casos como negativos (ausência de TMC). Houve 4% falsos positivos, ou seja, casos previstos como positivos mas que eram negativos na realidade e 16% falsos negativos, indicando casos que foram erroneamente classificados como negativos. Já para o ponto de corte = 0,24 o modelo classificou corretamente 15% casos como positivos e 60% casos com negativos, no entanto, houve um aumento de falsos positivos e uma redução de falsos negativos.

Tabela 3.10: Medidas de Desempenho para Regressão Logística.

Medidas de Desempenho		
Ponto de Corte	**0,5**	**0,24**
Acurácia	0,80	0,74
Sensibilidade	0,31	0,65
Especificidade	0,94	0,77
Precisão	0,64	0,46
Valor Predito Negativo	0,82	0,88
Estatística F1	0,42	0,54

Notamos pela Tabela 3.10 que com o ponto de corte padrão de 0,5, o modelo exibe uma acurácia de 80%, indicando que aproximadamente 80% das previsões foram corretas. A sensibilidade, que reflete a capacidade do modelo em identificar verdadeiros positivos, é de 31%. A especificidade, indicando a habilidade do modelo em identificar verdadeiros negativos, é alta, atingindo 94% e a precisão, representando a proporção de positivos previstos corretamente, é de 64%.

Ao ajustar o ponto de corte para 0,24, a acurácia reduz para 74%. Contudo, a sensibilidade melhora para 65%, indicando uma capacidade aprimorada do modelo em identificar casos positivos. A especificidade, embora reduza para 77%, ainda mantém um nível considerável e a precisão diminui para 46%.

O valor predito negativo aumenta de 82% para 88%, indicando uma maior capacidade do modelo em identificar corretamente casos negativos com o ponto de corte ajustado. A estatística F1 melhora de 42% para 54% com o ponto de corte de 0,24.

Observamos que o ajuste da Regressão Logística utilizando as covariáveis selecionadas pelo métodos *Stepwise* obteve resultados muito semelhantes àqueles alcançados com o ajuste da Regressão Logística com todas as covariáveis. Vale ressaltar, que os diferentes ajustes da Regressão Logística se assemelharam muito aos ajustes da Árvore de Classificação e da Floresta Aleatória, em todos esses cenários o ponto de corte obtidos pelas métricas Média-G e J apresentou melhores resultados devido ao equilíbrio entre sensibilidade e especificidade.

Capítulo 4

Considerações Finais

Nesse trabalho, apresentamos uma descrição do conjunto de dados, o qual é constituído por 3373 indivíduos após a exclusão das observações faltantes. Esse conjunto inclui uma variável resposta binária e 23 covariáveis. Adicionalmente, revisamos três algoritmos de classificação binária, sendo eles, árvores de classificação, florestas aleatórias e regressão logística, além de apresentarmos medidas de desempenho que foram utilizadas a fim de comparar os resultados obtidos em cada um dos algoritmos aplicados.

Primeiramente, realizamos uma análise descritiva e exploratória do banco de dados, em relação a variável resposta construída através do SQR20 (*Self-Reporting Questionnaire*) e categorizada em dois níveis, apresentando os gráficos adequados para as variáveis qualitativas e quantitativas seguidos de sua interpretação.

Em seguida, ajustamos os métodos de classificação binária utilizados para estudar a prevalência de transtornos mentais comuns (TMC) e aplicamos diferentes pontos de corte obtidos através das métricas Média-G e J devido ao desbalanceamento observado na variável resposta, em que 77% dos indivíduos não apresentam TMC. Dessa forma, observamos que os diferentes métodos adotados tiveram resultados muito semelhantes e, em todos os cenários, houve uma melhoria quando utilizado o ponto de corte proposto pelas métricas Média-G e J, uma vez que este provocava um melhor equilíbrio entre sensibilidade e especificidade. Além disso, o novo ponto de corte resultou em um aumento da sensibilidade, isso implica que o modelo teve uma melhora na capacidade de identificar casos verdadeiros positivos entre aqueles que realmente possuem transtorno mental comum (TMC). Esse aprimoramento é significativo no contexto de saúde mental, pois a redução de casos falsos negativos pode significar uma detecção mais eficaz de pessoas que necessitam de avaliação, tratamento ou suporte.

Por fim, concluímos que os diferentes métodos ajustados não tiveram grande impacto nos resultados dos modelos, portanto a escolha entre esses dependerá das prioridades específicas. Levando como prioridade a interpretabilidade do modelo o método de florestas aleatórias seria descartado, uma vez que árvores de classificação e regressão logística são geralmente mais interpretáveis. Desse modo, é importante ressaltar que no caso de árvores de classificação as variáveis selecionadas para a classificação de TMC foram:

- F101: indica como o indivíduo classifica seu estado de saúde;
- probs15: indica se o indivíduo teve problemas de saúde nos últimos 15 dias;
- k205b_hrs: indica o tempo total (em horas) que o indivíduo passa sentado durante um dia de final de semana.

Enquanto que para a regressão logística utilizando o método de seleção *Stepwise* foram:

- Sexo: indica o sexo do indivíduo;
- F101: indica como o indivíduo classifica seu estado de saúde;
- K201a: indica se atualmente o indivíduo trabalha ou faz trabalho voluntário fora de sua casa;
- k205b_hrs: indica o tempo total (em horas) que o indivíduo gasta sentado durante um dia de de final de semana;
- Fumo2: indica em qual categoria de tabagismo o indivíduo se enquadra;
- Conjugal3: indica qual a categoria da situação conjugal do indivíduo;
- probs15: indica se o indivíduo teve problemas de saúde nos últimos 15 dias.

Sendo assim, notamos que todas as variáveis selecionadas pela árvore de classificação também foram identificadas na regressão logística utilizando o método *Stepwise*. No entanto, a regressão logística incluiu quatro variáveis adicionais: Sexo, K201_a, Fumo2 e Conjugal3. Destas, destacamos que a inclusão da variável Sexo no modelo é condizente, uma vez que a variável resposta foi categorizada com base em pontos de corte específicos que consideram sexo e idade. Dessa maneira, concluímos que a abordagem mais indicada seria empregar a regressão logística.

Referências Bibliográficas

Araújo, Á. C. e Neto, F. L. (2014). A nova classificação americana para os transtornos mentais–o dsm-5. *Revista brasileira de terapia comportamental e cognitiva*, **16**(1), 67–82.

Bastos, A. A., Nogueira, L. R., Neto, J. V., Fisberg, R. M., Yannakoulia, M. e Ribeiro, S. M. L. (2020). Association between the adherence to the mediterranean dietary pattern and common mental disorders among community-dwelling elders: 2015 health survey of são paulo, sp, brazil. *Journal of Affective Disorders*, **265**, 389–394.

Breiman, L. (2001). Random forests. *Machine learning*, **45**, 5–32.

da Silva Filho, A. S. (2010). Inferência em amostras pequenas: método bootstrap. *Revista de Ciências exatas e tecnologia*, **5**(5), 115–126.

de Jesus Mari, J. e Williams, P. (1986). A validity study of a psychiatric screening questionnaire (srq-20) in primary care in the city of sao paulo. *The British Journal of Psychiatry*, **148**(1), 23–26.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, **27**(8), 861–874.

Giolo, S. R. (2021). *Introdução à análise de dados categóricos com aplicações*. Editora Blucher.

Hastie, T., Tibshirani, R., Friedman, J. H. e Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki.

James, G., Witten, D., Hastie, T., Tibshirani, R. *et al.* (2013). *An introduction to statistical learning*, volume 112. Springer.

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rahman, T., Ghosh, A. K., Shuvo, M. e Rahman, M. M. (2015). Mental stress recognition using k-nearest neighbor (knn) classifier on eeg signals. Em *Int. Conf. Materials, Electronics & Information Engineering (ICMEIE)*, páginas 1–4.

Scazufca, M., Menezes, P. R., Vallada, H. e Araya, R. (2009). Validity of the self reporting questionnaire-20 in epidemiological studies with older adults: results from the sao paulo ageing & health study. *Social psychiatry and psychiatric epidemiology*, **44**, 247–254.

Apêndice A

Código Análise Descritiva

```
### Ler base
dados <- read.csv("C:\\Users\\giuli\\OneDrive\\Documentos\\UFSCar\\TG\\Banco de Dados\\Banco_Andressa_SQR(1).csv",
sep=';',dec='.', encoding = "Latin-1 " )

### Arrumando banco de dados

dados [dados == ""] <- NA
dados$F101 [dados$F101 == "9"] <- NA
dados$k201a [dados$k201a == "9"] <- NA
dados$k205a_hrs [dados$k205a_hrs == "99"] <- NA
dados$k205b_hrs [dados$k205b_hrs == "99"] <- NA
dados$k206_hrs [dados$k206_hrs == "99"] <- NA
dados$k207_hrs [dados$k207_hrs == "99"] <- NA
dados$k208_hrs [dados$k208_hrs == "99"] <- NA
dados$k209_hrs [dados$k209_hrs == "99"] <- NA
dados$k209_hrs [dados$k209_hrs == "99"] <- NA
dados$k210 [dados$k210 == "9"] <- NA
dados$L01 [dados$L01 == "9"] <- NA
dados$L03 [dados$L03 == "99"] <- NA
dados$L10A [dados$L10A == "9"] <- NA
dados$L11 [dados$L11 == "9"] <- NA
dados$L12 [dados$L12 == "99"] <- NA
dados$L13 [dados$L13 == "9"] <- NA
dados$L13 [dados$L13 == "3"] <- NA
dados$L31 [dados$L31 == "9"] <- NA
dados$p01 [dados$p01 == "9"] <- NA
dados$alcool [dados$alcool == "não respondeu"] <- NA
dados$raca_cor [dados$raca_cor == "não respondeu"] <- NA
dados$fumo2 [dados$fumo2 == "não respondeu"] <- NA
dados$escolaridade_ind3 [dados$escolaridade_ind3 == "não respondeu"] <- NA
dados$conjugal3 [dados$conjugal3 == "NS/NR"] <- NA
dados$trabalho [dados$trabalho == "NS/NR"] <- NA
dados$probs15 [dados$probs15 == "NS/NR"] <- NA
dados$avsaude [dados$avsaude == "NS/NR"] <- NA
```

```
### Ler o arquivo .dta
SQR <- import(file = "C:\\Users\\giuli\\OneDrive\\Documentos\\UFSCar\\TG\\Banco de Dados\\banco_ISA2015_SQR-20.dta",
setclass = "data.frame")

SQR <- SQR %>% select("id","SRQ_cat")

### Juntando bases

base <- dados %>% left_join(SQR, by = "id")

### Verificando quantidade de NA's em cada variável
sapply(base, function(x) sum(is.na(x)))

### Retirando NA da variável resposta

base <- base[!is.na(base$SRQ_cat),]

### Analisando NA após limpar a variável resposta

sapply(base, function(x) sum(is.na(x)))

### Retirando variavéis com muitos NA e variáveis com informações repetidas

base <- base[,-c(17,19,24,30,15,28)]

### Analisando NA após todas as tratativas

sapply(base, function(x) sum(is.na(x)))


### Análise Descritiva

## Frequencia relativa de individuos com TMC

y <- table(base$SRQ_cat)
y <- prop.table(y)
y <- as.data.frame(y)
colnames(y) <- c("SQR", "Frequência Relativa")

ggplot(y, aes(y=`Frequência Relativa`, x=SQR)) +
  geom_bar(stat="identity", fill= "#00A9FF") +
  theme_classic(base_size = 11)+
  labs(x = "TMC",
       y = "Frequência Relativa") +
  scale_x_discrete(labels = c("Ausência", "Presença"))



## Variáveis Quantitativas
```

```
base$SRQ_cat <- as.factor(base$SRQ_cat)

#idade

ggplot(base, aes(y = idade, x = SRQ_cat, fill = SRQ_cat)) +
  geom_boxplot(show.legend = F,outlier.colour="black", outlier.shape=16,
               outlier.size=1.5) +
  theme_classic(base_size = 11) +
  xlab("TMC") +
  ylab("Idade")+
  scale_x_discrete(labels = c("Ausência", "Presença"))

base_idade_0 <- base %>% select(idade, SRQ_cat) %>% filter(SRQ_cat == 0)
summary(base_idade_0$idade)

base_idade_1 <- base %>% select(idade, SRQ_cat) %>% filter(SRQ_cat == 1)
summary(base_idade_1$idade)

#k205a_hrs

ggplot(base, aes(y = k205a_hrs, x = SRQ_cat, fill = SRQ_cat)) +
  geom_boxplot(show.legend = F,outlier.colour="black", outlier.shape=16,
               outlier.size=1.5) +
  theme_classic(base_size = 11) +
  xlab("TMC") +
  ylab("Tempo total de horas sentado \ndurante um dia de semana")+
  scale_x_discrete(labels = c("Ausência", "Presença"))

base_k205a_hrs_0 <- base %>% select(k205a_hrs, SRQ_cat) %>% filter(SRQ_cat == 0)
summary(base_k205a_hrs_0$k205a_hrs)

base_k205a_hrs_1 <- base %>% select(k205a_hrs, SRQ_cat) %>% filter(SRQ_cat == 1)
summary(base_k205a_hrs_1$k205a_hrs)

#k205b_hrs

ggplot(base, aes(y = k205b_hrs, x = SRQ_cat, fill = SRQ_cat)) +
  geom_boxplot(show.legend = F,outlier.colour="black", outlier.shape=16,
               outlier.size=1.5) +
  theme_classic(base_size = 11) +
  xlab("TMC") +
  ylab("Total de horas sentado \ndurante um dia de final de semana")+
  scale_x_discrete(labels = c("Ausência", "Presença"))

base_k205b_hrs_0 <- base %>% select(k205b_hrs, SRQ_cat) %>% filter(SRQ_cat == 0)
summary(base_k205b_hrs_0$k205b_hrs)

base_k205b_hrs_1 <- base %>% select(k205b_hrs, SRQ_cat) %>% filter(SRQ_cat == 1)
summary(base_k205b_hrs_1$k205b_hrs)
```

```
#k206_hrs

ggplot(base, aes(y = k206_hrs, x = SRQ_cat, fill = SRQ_cat)) +
  geom_boxplot(show.legend = F,outlier.colour="black", outlier.shape=16,
               outlier.size=1.5) +
  theme_classic(base_size = 11) +
  xlab("TMC") +
  ylab("Total de horas assistindo TV \ndurante um dia de semana")+
  scale_x_discrete(labels = c("Ausência", "Presença"))

base_k206_hrs_0 <- base %>% select(k206_hrs, SRQ_cat) %>% filter(SRQ_cat == 0)
summary(base_k206_hrs_0$k206_hrs)

base_k206_hrs_1 <- base %>% select(k206_hrs, SRQ_cat) %>% filter(SRQ_cat == 1)
summary(base_k206_hrs_1$k206_hrs)

#k207_hrs

ggplot(base, aes(y = k207_hrs, x = SRQ_cat, fill = SRQ_cat)) +
  geom_boxplot(show.legend = F,outlier.colour="black", outlier.shape=16,
               outlier.size=1.5) +
  theme_classic(base_size = 11) +
  xlab("TMC") +
  ylab("Total de horas assistindo TV \ndurante um dia de final de semana")+
  scale_x_discrete(labels = c("Ausência", "Presença"))

base_k207_hrs_0 <- base %>% select(k207_hrs, SRQ_cat) %>% filter(SRQ_cat == 0)
summary(base_k207_hrs_0$k207_hrs)

base_k207_hrs_1 <- base %>% select(k207_hrs, SRQ_cat) %>% filter(SRQ_cat == 1)
summary(base_k207_hrs_1$k207_hrs)

#k208_hrs

ggplot(base, aes(y = k208_hrs, x = SRQ_cat, fill = SRQ_cat)) +
  geom_boxplot(show.legend = F,outlier.colour="black", outlier.shape=16,
               outlier.size=1.5) +
  theme_classic(base_size = 11) +
  xlab("TMC") +
  ylab("Total de horas no computador \ndurante um dia de semana")+
  scale_x_discrete(labels = c("Ausência", "Presença"))

base_k208_hrs_0 <- base %>% select(k208_hrs, SRQ_cat) %>% filter(SRQ_cat == 0)
summary(base_k208_hrs_0$k208_hrs)

base_k208_hrs_1 <- base %>% select(k208_hrs, SRQ_cat) %>% filter(SRQ_cat == 1)
summary(base_k208_hrs_1$k208_hrs)

#k209_hrs
```

```
ggplot(base, aes(y = k209_hrs, x = SRQ_cat, fill = SRQ_cat)) +
  geom_boxplot(show.legend = F,outlier.colour="black", outlier.shape=16,
               outlier.size=1.5) +
  theme_classic(base_size = 11) +
  xlab("TMC") +
  ylab("Total de horas no computador \ndurante um dia de final de semana")+
  scale_x_discrete(labels = c("Ausência", "Presença"))

base_k209_hrs_0 <- base %>% select(k209_hrs, SRQ_cat) %>% filter(SRQ_cat == 0)
summary(base_k209_hrs_0$k209_hrs)

base_k209_hrs_1 <- base %>% select(k209_hrs, SRQ_cat) %>% filter(SRQ_cat == 1)
summary(base_k209_hrs_1$k209_hrs)

## Variáveis Qualitativas

#sexo

contagem_sexo <- table(base$SRQ_cat,base$sexo)
porcentagem_sexo <- prop.table(contagem_sexo, margin = 1)
porcentagem_sexo <- as.data.frame.matrix(porcentagem_sexo)
porcentagem_sexo$SRQ_cat <- rownames(porcentagem_sexo)
porcentagem_sexo <- reshape2::melt(porcentagem_sexo, id.vars = "SRQ_cat")



ggplot(porcentagem_sexo, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Sexo") +
  scale_x_discrete(labels = c("Ausência", "Presença"))+
  scale_fill_discrete(labels = c("Feminino", "Masculino"))

#F101

contagem_F101 <- table(base$SRQ_cat,base$F101)
porcentagem_F101 <- prop.table(contagem_F101, margin = 1)
porcentagem_F101 <- as.data.frame.matrix(porcentagem_F101)
porcentagem_F101$SRQ_cat <- rownames(porcentagem_F101)
porcentagem_F101 <- reshape2::melt(porcentagem_F101, id.vars = "SRQ_cat")

ggplot(porcentagem_F101, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Estado de saúde declarado \npelo indivíduo") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Excelente/Muito boa", "Boa", "Regular",
```

```
                                "Ruim", "Muito ruim"))



#k201a

contagem_k201a <- table(base$SRQ_cat,base$k201a)
porcentagem_k201a <- prop.table(contagem_k201a, margin = 1)
porcentagem_k201a <- as.data.frame.matrix(porcentagem_k201a)
porcentagem_k201a$SRQ_cat <- rownames(porcentagem_k201a)
porcentagem_k201a <- reshape2::melt(porcentagem_k201a, id.vars = "SRQ_cat")

ggplot(porcentagem_k201a, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Trabalha ou faz trabalho \nvoluntário fora de casa?") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Não", "Sim"))



#k210

contagem_k210 <- table(base$SRQ_cat,base$k210)
porcentagem_k210 <- prop.table(contagem_k210, margin = 1)
porcentagem_k210 <- as.data.frame.matrix(porcentagem_k210)
porcentagem_k210$SRQ_cat <- rownames(porcentagem_k210)
porcentagem_k210 <- reshape2::melt(porcentagem_k210, id.vars = "SRQ_cat")

ggplot(porcentagem_k210, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Pratica regularmente algum \ntipo de exercício físico ou esporte?") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Não", "Sim"))



#L01

contagem_L01 <- table(base$SRQ_cat,base$L01)
porcentagem_L01 <- prop.table(contagem_L01, margin = 1)
porcentagem_L01 <- as.data.frame.matrix(porcentagem_L01)
porcentagem_L01$SRQ_cat <- rownames(porcentagem_L01)
porcentagem_L01 <- reshape2::melt(porcentagem_L01, id.vars = "SRQ_cat")

ggplot(porcentagem_L01, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
```

```
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Cor/Raça") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Branca", "Preta", "Amarela","Parda",
                                 "Indígena", "Outra"))

#L03

contagem_L03 <- table(base$SRQ_cat,base$L03)
porcentagem_L03 <- prop.table(contagem_L03, margin = 1)
porcentagem_L03 <- as.data.frame.matrix(porcentagem_L03)
porcentagem_L03$SRQ_cat <- rownames(porcentagem_L03)
porcentagem_L03 <- reshape2::melt(porcentagem_L03, id.vars = "SRQ_cat")

ggplot(porcentagem_L03, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Religião") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Nenhuma", "Evangélica/Protestante", "Católica","Espírita",
                                 "Judaísmo", "Budismo", "Umbanda/ Candomblé", "Islamismo",
                                 "Outras"))


#L11

contagem_L11 <- table(base$SRQ_cat,base$L11)
porcentagem_L11 <- prop.table(contagem_L11, margin = 1)
porcentagem_L11 <- as.data.frame.matrix(porcentagem_L11)
porcentagem_L11$SRQ_cat <- rownames(porcentagem_L11)
porcentagem_L11 <- reshape2::melt(porcentagem_L11, id.vars = "SRQ_cat")

ggplot(porcentagem_L11, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Tem filhos?") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Não", "Sim"))

#L13

contagem_L13 <- table(base$SRQ_cat,base$L13)
porcentagem_L13 <- prop.table(contagem_L13, margin = 1)
porcentagem_L13 <- as.data.frame.matrix(porcentagem_L13)
```

```
porcentagem_L13$SRQ_cat <- rownames(porcentagem_L13)
porcentagem_L13 <- reshape2::melt(porcentagem_L13, id.vars = "SRQ_cat")

ggplot(porcentagem_L13, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Frequenta atualmente \nalgum curso regular?") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Não", "Sim"))

#p01

contagem_p01 <- table(base$SRQ_cat,base$p01)
porcentagem_p01 <- prop.table(contagem_p01, margin = 1)
porcentagem_p01 <- as.data.frame.matrix(porcentagem_p01)
porcentagem_p01$SRQ_cat <- rownames(porcentagem_p01)
porcentagem_p01 <- reshape2::melt(porcentagem_p01, id.vars = "SRQ_cat")

ggplot(porcentagem_p01, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Possui algum animal \nem seu domicílio?") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Não", "Sim"))

#imc_cat

contagem_imc_cat <- table(base$SRQ_cat,base$imc_cat)
porcentagem_imc_cat <- prop.table(contagem_imc_cat, margin = 1)
porcentagem_imc_cat <- as.data.frame.matrix(porcentagem_imc_cat)
porcentagem_imc_cat$SRQ_cat <- rownames(porcentagem_imc_cat)
porcentagem_imc_cat <- reshape2::melt(porcentagem_imc_cat, id.vars = "SRQ_cat")

ggplot(porcentagem_imc_cat, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "IMC") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Baixo peso", "Eutrofia", "Obesidade", "Sobrepeso"))

#fumo2

contagem_fumo2 <- table(base$SRQ_cat,base$fumo2)
porcentagem_fumo2 <- prop.table(contagem_fumo2, margin = 1)
```

```
porcentagem_fumo2 <- as.data.frame.matrix(porcentagem_fumo2)
porcentagem_fumo2$SRQ_cat <- rownames(porcentagem_fumo2)
porcentagem_fumo2 <- reshape2::melt(porcentagem_fumo2, id.vars = "SRQ_cat")

ggplot(porcentagem_fumo2, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Categorias de tabagismo") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Ex-fumante e fumante","Nunca fumou"))



#alcool

contagem_alcool <- table(base$SRQ_cat,base$alcool)
porcentagem_alcool <- prop.table(contagem_alcool, margin = 1)
porcentagem_alcool <- as.data.frame.matrix(porcentagem_alcool)
porcentagem_alcool$SRQ_cat <- rownames(porcentagem_alcool)
porcentagem_alcool <- reshape2::melt(porcentagem_alcool, id.vars = "SRQ_cat")

ggplot(porcentagem_alcool, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Classificação do consumo de \nbebida alcoólica") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Bebe atualmente","Nunca bebeu", "Parou de beber"))



#escolaridade_ind3

contagem_escolaridade_ind3 <- table(base$SRQ_cat,base$escolaridade_ind3)
porcentagem_escolaridade_ind3 <- prop.table(contagem_escolaridade_ind3, margin = 1)
porcentagem_escolaridade_ind3 <- as.data.frame.matrix(porcentagem_escolaridade_ind3)
porcentagem_escolaridade_ind3$SRQ_cat <- rownames(porcentagem_escolaridade_ind3)
porcentagem_escolaridade_ind3 <- reshape2::melt(porcentagem_escolaridade_ind3, id.vars = "SRQ_cat")

ggplot(porcentagem_escolaridade_ind3, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Escolaridade") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Maior 12 anos", "10 a 12 anos",
                                 "Menor/Igual 9 anos"))
```

```
#AF_GLOBAL_cat_OMS_1

contagem_AF_GLOBAL_cat_OMS_1 <- table(base$SRQ_cat,base$AF_GLOBAL_cat_OMS_1)
porcentagem_AF_GLOBAL_cat_OMS_1 <- prop.table(contagem_AF_GLOBAL_cat_OMS_1, margin = 1)
porcentagem_AF_GLOBAL_cat_OMS_1 <- as.data.frame.matrix(porcentagem_AF_GLOBAL_cat_OMS_1)
porcentagem_AF_GLOBAL_cat_OMS_1$SRQ_cat <- rownames(porcentagem_AF_GLOBAL_cat_OMS_1)
porcentagem_AF_GLOBAL_cat_OMS_1 <- reshape2::melt(porcentagem_AF_GLOBAL_cat_OMS_1, id.vars = "SRQ_cat")

ggplot(porcentagem_AF_GLOBAL_cat_OMS_1, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Nível de atividade física") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Cumpre a recomendação",
                                 "Não cumpre a recomendação"))


#conjugal3

contagem_conjugal3 <- table(base$SRQ_cat,base$conjugal3)
porcentagem_conjugal3 <- prop.table(contagem_conjugal3, margin = 1)
porcentagem_conjugal3 <- as.data.frame.matrix(porcentagem_conjugal3)
porcentagem_conjugal3$SRQ_cat <- rownames(porcentagem_conjugal3)
porcentagem_conjugal3 <- reshape2::melt(porcentagem_conjugal3, id.vars = "SRQ_cat")

ggplot(porcentagem_conjugal3, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
  labs(x = "TMC",
       y = "Frequência Relativa",
       fill = "Situação Conjugal") +
  scale_x_discrete(labels = c("Ausência", "Presença")) +
  scale_fill_discrete(labels = c("Casado/União Estável","Separado/Desquitado/Divorciado",
                                 "Solteiro", "Viúvo"))


#probs15

contagem_probs15 <- table(base$SRQ_cat,base$probs15)
porcentagem_probs15 <- prop.table(contagem_probs15, margin = 1)
porcentagem_probs15 <- as.data.frame.matrix(porcentagem_probs15)
porcentagem_probs15$SRQ_cat <- rownames(porcentagem_probs15)
porcentagem_probs15 <- reshape2::melt(porcentagem_probs15, id.vars = "SRQ_cat")

ggplot(porcentagem_probs15, aes(x = SRQ_cat, y = value, fill = variable)) +
  geom_col(position = "dodge") +
  theme_classic(base_size = 11) +
```

```
labs(x = "TMC",
     y = "Frequência Relativa",
     fill = "Problema de saúde \nnos últimos 15 dias") +
scale_x_discrete(labels = c("Ausência", "Presença")) +
scale_fill_discrete(labels = c("Não","Sim"))
```

Apêndice B

Código Divisão do Banco de Dados

```
df <- base
attach(df)

#Divida o conjunto de dados em 70% para treinamento e 30% para teste.

set.seed(20230625)  # Definir uma semente para reprodução dos resultados
indices <- createDataPartition(y = df$SRQ_cat, p = 0.7, list = FALSE)
dados_treino <- df[indices, ]  # 70% dos dados para treinamento
dados_teste <- df[-indices, ]  # 30% dos dados para teste
```

Apêndice C

Código Árvore de Classificação

```
library(rpart); library(rpart.plot)

fit <- rpart(Y_treino ~ ., data = dados_treino[, -24], method = "class")
melhor_cp <- fit$cptable[which.min(fit$cptable[, "xerror"]),"CP"]
pfit <- rpart::prune(fit, cp = melhor_cp)
rpart.plot(pfit, type = 5)
preditos_arvore <- predict(pfit, dados_teste[,-24])

summary(pfit)

#Curva ROC
roc_curve <- roc(response=y_teste, predictor = preditos_arvore[,2])

# Plotar a curva ROC
plot(roc_curve, main = "Curva ROC", col = "cornflowerblue", lwd = 2, xlab = "Especificidade", ylab = "Sensibilidade")
auc_value <- auc(roc_curve)
text(0.5, 0.2, paste("AUC =", round(auc_value, 2)), col = "cornflowerblue", cex = 1.2)

g_mean <- sqrt(roc_curve$sensitivities*roc_curve$specificities)
max_g_mean<- max(g_mean)
id <- which(g_mean==max_g_mean)
g_final <- roc_curve$thresholds[id]
g_final

j_estat <- roc_curve$sensitivities+roc_curve$specificities-1
max_j_estat<- max(j_estat)
id <- which(j_estat==max_j_estat)
j_final <- roc_curve$thresholds[id]
j_final

#Matriz de Confusão

previsoes_binarias <- ifelse(preditos_arvore[,2] > j_final, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias),
                reference = as.factor(y_teste),
```

```
                positive = '1',
                mode = "everything")

previsoes_binarias2 <- ifelse(preditos_arvore[,2] > 0.5, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias2),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")
```

Apêndice D

Código Florestas Aleatórias

```
library(randomForest);library(ranger)

floresta <- randomForest(Y_treino ~ ., data = dados_treino[, -24], method = "class", probability = TRUE)

varImpPlot(floresta)

preditos_floresta <- predict(floresta, dados_teste[,-24], type = "prob")

#Curva ROC
roc_curve <- roc(response=y_teste, predictor = preditos_floresta[,2])

# Plotar a curva ROC
plot(roc_curve, main = "Curva ROC", col = "cornflowerblue", lwd = 2, xlab = "Especificidade", ylab = "Sensibilidade")
auc_value <- auc(roc_curve)
text(0.5, 0.2, paste("AUC =", round(auc_value, 2)), col = "cornflowerblue", cex = 1.2)

g_mean <- sqrt(roc_curve$sensitivities*roc_curve$specificities)
max_g_mean<- max(g_mean)
id <- which(g_mean==max_g_mean)
g_final <- roc_curve$thresholds[id]
g_final

j_estat <- roc_curve$sensitivities+roc_curve$specificities-1
max_j_estat<- max(j_estat)
id <- which(j_estat==max_j_estat)
j_final <- roc_curve$thresholds[id]
j_final

#Matriz de Confusão

previsoes_binarias <- ifelse(preditos_floresta[,2] > j_final, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")
```

```
previsoes_binarias2 <- ifelse(preditos_floresta[,2] > 0.5, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias2),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")
```

Apêndice E

Código Regressão Logística

```
Y_treino = dados_treino$SRQ_cat

logistic.model <- glm(Y_treino ~ .,
                      data = dados_treino[,-24],
                      family = binomial(link = "logit"))


summary(logistic.model)

#coeficientes do modelo
odd.ratio = exp(coef(logistic.model))

#previsão no teste
previsoes <- predict(logistic.model, newdata = dados_teste[,-24], type= "response")

y_teste <- unlist(dados_teste[,24])

#Curva ROC
roc_curve <- roc(response= y_teste, predictor= previsoes)

# Plotar a curva ROC
plot(roc_curve, main = "Curva ROC", col = "cornflowerblue", lwd = 2)
auc_value <- auc(roc_curve)
text(0.5, 0.2, paste("AUC =", round(auc_value, 2)), col = "cornflowerblue", cex = 1.2)

g_mean <- sqrt(roc_curve$sensitivities*roc_curve$specificities)
max_g_mean<- max(g_mean)
id <- which(g_mean==max_g_mean)
g_final <- roc_curve$thresholds[id]
g_final

j_estat <- roc_curve$sensitivities+roc_curve$specificities-1
max_j_estat<- max(j_estat)
id <- which(j_estat==max_j_estat)
j_final <- roc_curve$thresholds[id]
```

```
j_final

#Matriz de Confusão

previsoes_binarias <- ifelse(previsoes > j_final, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")


previsoes_binarias2 <- ifelse(previsoes > 0.5, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias2),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")

###retirando variaveis nao significativas

step(logistic.model, direction = "both")

logistic.model_ajust<- glm(Y_treino ~ sexo + F101 + k201a + k205b_hrs +
                           fumo2 + conjugal3 + probs15, data= dados_treino[,-24],
                           family = binomial(link = "logit"))


summary(logistic.model_ajust)


#coeficientes do modelo
odd.ratio = exp(coef(logistic.model_ajust))

#previs?o no teste
previsoes <- predict(logistic.model_ajust, newdata = dados_teste[,-24], type= "response")

y_teste <- unlist(dados_teste[,24])

#Curva ROC
roc_curve <- roc(response= y_teste, predictor= previsoes)

# Plotar a curva ROC
plot(roc_curve, main = "Curva ROC", col = "cornflowerblue", lwd = 2, xlab = "Especificidade", ylab = "Sensibilidade")
auc_value <- auc(roc_curve)
text(0.5, 0.2, paste("AUC =", round(auc_value, 2)), col = "cornflowerblue", cex = 1.2)

g_mean <- sqrt(roc_curve$sensitivities*roc_curve$specificities)
max_g_mean<- max(g_mean)
id <- which(g_mean==max_g_mean)
g_final <- roc_curve$thresholds[id]
g_final
```

```
j_estat <- roc_curve$sensitivities+roc_curve$specificities-1
max_j_estat<- max(j_estat)
id <- which(j_estat==max_j_estat)
j_final <- roc_curve$thresholds[id]
j_final

#Matriz de Confusão

previsoes_binarias <- ifelse(previsoes > j_final, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")


previsoes_binarias2 <- ifelse(previsoes > 0.5, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias2),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")
```