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

Utilização de algoritmos de classificação para diagnóstico de diabetes

Larissa de Oliveira

Trabalho de Conclusão de Curso

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

Utilização de algoritmos de classificação para diagnóstico de diabetes

Larissa de Oliveira
Orientadora: Profª Drª Andressa Cerqueira

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

São Carlos
Fevereiro de 2024

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

Application of classification algorithms for the diagnosis of diabetes

Larissa de Oliveira
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

Larissa de Oliveira

Utilização de algoritmos de classificação para diagnóstico de diabetes

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido pela Larissa de Oliveira e aprovado pela banca examinadora.

Aprovado em 23 de Janeiro de 2024

Banca Examinadora:

- Andressa Cerqueira (Orientadora)
- Maria Sílvia de Assis Moura
- Michel Helcias Montoril

Resumo

O presente Trabalho de Graduação consiste no estudo de algoritmos de classificação binária para diagnóstico de diabetes com base em dados clínicos, dados de atividades físicas e dados de entretenimento tecnológico. Para esse objetivo, realizaremos uma definição estatística e conceitual do problema em estudo e obteremos os resultados a partir dos métodos de Regressão Logística e Árvores de Classificação. Todas as aplicações computacionais serão feitas através da linguagem de programação R.

Palavras-chave: *diabetes, classificação, Regressão Logística e Árvores de Classificação.*

Abstract

The present undergraduate thesis consists of the study of binary classification algorithms for diagnosing diabetes based on clinical data and physical activity data. For this purpose, we present a statistical and conceptual definition of the problem and obtain the results from the methods of Logistic Regression and Classification Trees. All computational applications will be done using the R programming language.

Keywords: *diabetes, classification, Logistic Regression and Classification Trees.*

Sumário

Capítulo 1

Introdução

A diabetes é uma doença crônica causada pela produção insuficiente ou má absorção de insulina, o que resulta em níveis elevados de glicose no sangue. Caso a diabetes não seja tratada ou identificada, muitas complicações podem ocorrer, portanto é essencial o processo de identificação da doença. Desse modo, utilizando recursos estatísticos, dados clínicos, dados de atividade física e dados de entretenimento tecnológico, podemos predizer esse processo de identificação da doença de forma simplificada.

Métodos de classificação são métodos estatísticos muito utilizados em diversas áreas de pesquisa, como, por exemplo, na área da saúde. Esses métodos têm o objetivo de fornecer modelos conhecidos como classificadores, que classificam dados em categorias distintas com um alto poder preditivo.

Em Sisodia e Sisodia (2018), há um exemplo de aplicação de Árvores de Classificação no estudo sobre a predição de diabetes em pacientes. Nessa abordagem, criada em 1984, temos diversas repartições denotadas por "nós" na árvore que verificam se uma condição, descrita no "nó" acima, é verdadeira ou não, assim o processo continua até obter o resultado de diagnóstico de diabetes dos pacientes.

Outra abordagem é aplicada por Hassan e Amiri (2019), em que há a predição de diabetes utilizando um banco de dados desbalanceado, ou seja, no banco de dados há uma proporção muito maior de indivíduos não diabéticos do que indivíduos que possuem a doença. Entre os classificadores utilizados no estudo, temos a Regressão Logística, que permite o uso um método de regressão para predizer a probabilidade de um evento, como também utiliza-se do método de Máxima Verossimilhança para a estimação dos coeficientes da regressão.

Izbicki e dos Santos (2020) apresentam uma ampla abordagem dos métodos de Re-

gressão Logística e Árvores de Classificação, além de um estudo abrangente sobre desbalanceamento de classes.

O objetivo do presente trabalho de conclusão de curso é a classificação binária do diagnóstico de diabetes em pacientes com base em dados clínicos, dados de práticas de atividades físicas e dados de entretenimento tecnológico. Com essa finalidade, realizamos a classificação a partir dos métodos de Regressão Logística e Árvores de Classificação.

Desse modo, primeiramente, apresentamos os métodos de forma técnica e conceitual, suas respectivas definições, aplicações e direcionamos cada um deles no nosso principal objetivo, ou seja, na classificação do diagnóstico de diabetes a partir dos dados disponíveis.

Após a obtenção dos resultados, utilizando todos os testes, temos predições sobre o diagnóstico de diabetes de pacientes, a partir das covariáveis disponibilizadas no banco de dados. Finalmente, comparamos os métodos utilizados, a fim de buscar o método que apresenta um melhor resultado de predição.

Esse Trabalho de Conclusão de Curso está dividido em 5 capítulos, sendo o primeiro capítulo o atual capítulo, em que há a introdução do trabalho. No segundo capítulo , é apresentada a revisão dos métodos de classificação utilizados: Regressão Logística e Árvores de Classificação. Abordamos as definições, os principais aspectos, vantagens e desvantagens de cada um dos métodos. Ademais, são apresentadas as medidas de desempenho que são utilizadas para compararmos a performance de cada um dos métodos, o método de divisão do banco de dados e os métodos para tratar o desbalanceamento de classes. No terceiro capítulo é apresentada a descrição e estruturação do banco de dados, além da análise descritiva e exploratória do conjunto de variáveis utilizadas no estudo do diagnóstico de diabetes. No quarto capítulo é apresentada a aplicação dos métodos de classificação no banco de dados disponibilizado para a realização do trabalho. No quinto capítulo são relatadas as considerações finais deste trabalho, assim como suas contribuições. Por fim, no apêndice A foi disponibilizado o código utilizado para a elaboração do trabalho, construído em linguagem de programação R R Development Core Team (2020).

Capítulo 2

Revisão da Literatura

Nesse capítulo é abordada a revisão de toda a literatura utilizada para a realização do trabalho de graduação, como os métodos de classificação, sendo eles: Regressão Logística e Árvore de Classificação, além das medidas de desempenho, o método de divisão do banco de dados e os métodos de tratar o desbalanceamento de classes.

2.1 Regressão Logística

Primeiramente, assumimos que a variável resposta Y é binária, ou seja, ela pode assumir duas categorias, 0 ou 1, $Y \in \{0, 1\}$. Nesse trabalho, Y é definido da seguinte forma:

$$Y = \begin{cases} 1, \text{ se o indivíduo é diabético;} \\ 0, \text{ se o indivíduo não é diabético.} \end{cases}$$

É de nosso interesse estimar a probabilidade de um indivíduo ser diabético segundo a informação de um conjunto de variáveis explicativas que são, geralmente, uma mistura de variáveis categóricas e quantitativas, denotadas por $\boldsymbol{x} = (x_1, ..., x_d)$, ou seja, desejamos calcular $\mathbb{P}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta})$, em que $\boldsymbol{\beta}$ representa os parâmetros de distribuição. Para isso, se fosse considerado um modelo de regressão linear múltipla para representar essa probabilidade, ela seria expressa da seguinte forma:

$$\mathbb{P}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta}) = \mathbb{E}[Y|\boldsymbol{x}, \boldsymbol{\beta}] = \beta_0 + \sum_{i=1}^{d} \beta_i x_i,$$

em que β_0, β_i são os coeficientes de uma regressão linear múltipla e $i = 1, ..., d$.

Entretanto, a abordagem considerando um modelo de regressão linear múltipla pode gerar resultados inconsistentes como estimativas para $\mathbb{P}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta})$ menores do que 0 ou maiores do que 1. Para evitar esse problema, consideramos utilizar a Regressão Logística, que fornece resultados entre 0 e 1 para todos os valores de $\boldsymbol{x}$. Além disso, a relação entre $\boldsymbol{x}$ e $\mathbb{E}[Y|\boldsymbol{x}, \boldsymbol{\beta}]$ tem padrão sigmoidal.

A Regressão Logística é um método de classificação que possui a finalidade de predizer a probabilidade de um evento específico, ou seja, uma variável resposta qualitativa que pode ser expressa por duas ou mais categorias.

A variável resposta pode assumir valores nominais ou ordinais, por isso é necessário utilizar uma abordagem adequada para a análise dessas variáveis. Em relação as categorias ordinais, existe uma ordem entre elas e para modelarmos esse caso, utilizamos a Regressão Logística Ordinal. Nesse trabalho de conclusão de curso, a variável em estudo é uma variável nominal e com isso utilizamos a Regressão Logística Nominal, em que não há ordem entre as categorias da variável independente.

Desse modo, temos a seguinte forma paramétrica da Regressão Logística:

$$\mathbb{P}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta}) = \frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}. \tag{2.1}$$

Consequentemente, a probabilidade complementar de $\mathbb{P}(Y = 1|\boldsymbol{x})$ é dada por:

$$\mathbb{P}(Y = 0|\boldsymbol{x}, \boldsymbol{\beta}) = \frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}. \tag{2.2}$$

Como apresentado por Giolo (2021), após algumas manipulações matemáticas na forma paramétrica da regressão logística (2.1), obtemos o logaritmo da chance (*log odds* ou *logit*), dada por:

$$\log\left(\frac{\mathbb{P}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta})}{\mathbb{P}(Y = 0|\boldsymbol{x}, \boldsymbol{\beta})}\right) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i. \tag{2.3}$$

Em relação ao estudo do diagnóstico da diabetes, o logaritmo da chance é o logaritmo da probabilidade de um indivíduo possuir diabetes dado o conjunto de covariáveis, dividida pela probabilidade do indivíduo não possuir diabetes dado o conjunto de covariáveis.

Apesar da forma paramétrica (2.1) não ser linear, ela é equivalente a (2.3), na qual existe uma relação linear entre as covariáveis e a variável resposta. Além disso, (2.3)

tem a propriedade de que os resultados obtidos em $\beta_0 + \sum_{i=1}^{d} \beta_i x_i$ possuem um valor correspondente entre 0 e 1 para $\mathbb{P}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta})$. Portanto, as probabilidades estimadas pelo logaritmo da chance são valores pertencentes ao intervalo $[0, 1]$.

Segundo Izbicki e dos Santos (2020), para a estimação dos coeficientes do modelo de Regressão Logística, utilizamos o Método de Máxima Verossimilhança. Para isso, é considerada uma amostra i.i.d. (independente e identicamente distribuída), ou seja, uma amostra coletada de forma independente e seguindo a mesma distribuição de probabilidade, $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$, contendo n observações, onde $\boldsymbol{x}_k = (x_{k,1}, ..., x_{k,d})$. Desta maneira, a função de verossimilhança condicional nas covariáveis é:

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

em que $i = 1, ..., d$ e $k = 1, ..., n$.

Em seguida, aplicando-se o logaritmo da função de verossimilhança, tem-se a equação:

$$l(y; \boldsymbol{x}, \boldsymbol{\beta}) = \sum_{k=1}^{n} \left[y_k \log \left(\frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{k,i}}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{k,i}}} \right) + (1 - y_k) \log \left(\frac{1}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_{k,i}}} \right) \right],$$

em que as estimativas de $\beta_0, \beta_1, ..., \beta_p$ são aquelas que maximizam $l(y; \boldsymbol{x}, \boldsymbol{\beta})$.

Para a finalidade de maximizar o logaritmo da função de verossimilhança, essa função é derivada com respeito a cada um dos coeficientes e igualada a 0. Após isso, realizamos a segunda derivada, para garantirmos que os coeficientes maximizam a função, ou seja, a segunda derivada com respeito a cada um dos coeficientes deve ser menor do que 0. Para essa maximização, fazem-se necessário utilizar métodos numéricos como o método de *Newton-Raphson*.

O uso de Regressão Logística para problemas de classificação possui algumas vantagens como a interpretabilidade, pois os coeficientes da regressão podem ser facilmente interpretados como o efeito que cada variável tem sobre a probabilidade da classe de interesse acontecer. Além disso, é um método computacionalmente eficiente, pois é de fácil e flexível implementação computacional, especialmente em comparação com algoritmos mais complexos.

Na Regressão Logística pode existir um número grande de covariáveis e muitas vezes

nem todas são relevantes para o estudo. Desse modo, um método eficiente de seleção de variáveis é o *Stepwise*. As técnicas do método *Stepwise* são: *Forward Selection*, *Backward Elimination* e a Bidirecional que é uma combinação das técnicas anteriores.

O método *Stepwise*, utilizando a técnica *Forward Selection*, começa com um modelo vazio e adiciona uma covariável por vez, escolhendo aquela que oferece a maior melhoria no ajuste do modelo. O processo acaba quando não há mais variáveis que contribuam significativamente. Por outro lado, o método *Stepwise* utilizando a técnica *Backward Elimination* inicia com um modelo que inclui todas as variáveis e remove uma por vez, escolhendo aquela que menos contribui para o ajuste. O processo finaliza até que a remoção de variáveis não melhore significativamente o modelo.

Além desses, o método *Stepwise* utilizando a técnica Bidirecional alterna entre adição e remoção de variáveis, escolhendo aquela covariável que proporciona a maior melhoria modelo.

O critério de informação AIC (*Akaike Information Criterion*) será o critério utilizado para medir a melhoria modelo ao utilizar o método *Stepwise*. O AIC busca encontrar um equilíbrio entre a adequação do modelo aos dados e a complexidade do próprio modelo.

O método *Stepwise* é altamente eficiente na escolha de um conjunto otimizado de covariáveis. Além disso, temos maior interpretabilidade, pois modelos com menos variáveis facilitam a compreensão das relações entre elas. E também, a economia computacional é uma consideração relevante, uma vez que a redução no número de variáveis torna a execução do modelo mais rápida e eficiente.

2.2 Árvore de Classificação

A utilização de árvores de decisão para problemas de classificação é uma abordagem muito popular na área de estatística e aprendizado de máquina.

Uma árvore de classificação tem a finalidade de classificar a variável resposta, ajustando um modelo simples baseado em particionamentos recursivos no espaço das variáveis independentes.

Em uma árvore, denomina-se nó cada particionamento da mesma. Desse modo, o nó inicial é dado pela amostra inicial, nós intermediários são dados por subamostras que geraram novas subamostras e por fim cada nó final tem-se o nome de folha. A Figura 2.1, disponibilizada abaixo, apresenta os termos mencionados numa árvore de classificação.

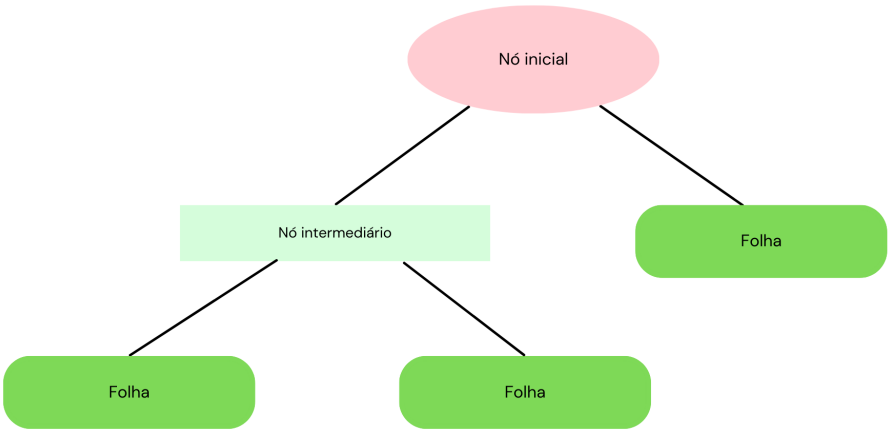


Figura 2.1: Exemplo de estrutura de uma árvore de classificação.

Para a construção de uma árvore para prever novas observações, segue-se alguns passos recursivos. Primeiramente, na amostra inicial, verifica-se se a condição descrita no primeiro nó é verdadeira ou não. Caso seja verdadeira, seguimos para a esquerda. Caso contrário, seguimos para a direita. O processo continua verificando se as condições descritas nos próximos nós são verdadeiras até atingir uma folha. Denominamos por $R_1, ..., R_j$ as partições da árvore em regiões distintas e disjuntas.

Segundo Izbicki e dos Santos (2020), a predição da variável resposta Y em uma região constante R_m é dada por:

$$g(\boldsymbol{x}) = \text{moda}\{y_i : \boldsymbol{x}_i \in R_m\}$$

para $i = 1, ..., n$.

Desta maneira, para predizer Y, observamos a região em que as variáveis independentes se encontram e calculamos a moda das variáveis respostas de $\boldsymbol{x}$ pertencentes ao conjunto de treinamento correspondente a essa região.

Ademais, a construção de uma árvore de classificação é dividida em dois passos, sendo o primeiro a criação de uma árvore complexa. No andamento do primeiro passo, pode-se considerar alguns critérios para encontrar as melhores partições no processo de construção da árvore, sendo eles:

- Índice de Gini: $\sum_R \sum_{c \in \mathcal{C}} \hat{p}_{R,c}(1 - \hat{p}_{R,c})$,
- Erro de Classificação (*Misclassification error*): $1 - \hat{p}_{R,c(R)}$,

- Desvio (*Deviance*): $-\sum_{c=1}^{\mathcal{C}} \hat{p}_{R,c} \ln(\hat{p}_{R,c})$,

em que denotamos por R as regiões da árvore. Além disso, $\hat{p}_{R,c}$ representa a proporção das observações presentes na região R que são classificadas como sendo da categoria c e, por fim, $c(R) = \arg\max_c \hat{p}_{R,c}$, ou seja, a classe mais frequente da região R.

Segundo Hastie *et al.* (2009), os três critérios apresentados são semelhantes, mas como o Índice de Gini e o *Deviance* são diferenciáveis, são mais passíveis de otimização numérica. Além disso, eles são mais sensíveis a mudanças nas probabilidades nas regiões.

Nesse trabalho de graduação optamos pela utilização do índice de Gini. Uma característica desse índice é que sua minimização ocorre quando todas as proporções $\hat{p}_{R,c}$ são iguais à 0 ou 1, assim tem-se que a melhor partição de todos os passos é a que contém apenas observações de uma mesma classe, ou seja, uma árvore pura. Desse modo, o ideal é utilizar a partição que minimiza esse índice.

Portanto, construímos uma árvore "grande" com muitas repartições selecionando as variáveis que melhor separam os dados em relação às classes, ou seja, são utilizadas as partições que possuem menor índice de Gini. Assim sendo, a árvore cresce de maneira recursiva até atingir uma folha, ou seja, até atingir um critério de parada.

O segundo passo na construção de uma árvore de classificação é a etapa de podagem, que consiste na remoção de nós pouco significativos para melhorar a generalização do modelo e evitar o *overfitting*, ou seja, o superajuste ao conjunto de dados.

Por fim, após esses passos, tem-se a estrutura de uma árvore de classificação que pode classificar a variável resposta a partir de informações das variáveis independentes.

Existem diversas vantagens em utilizar árvores de classificação, como por exemplo a facilidade de seleção de covariáveis, pois ela é realizada de forma automática, ou seja, quando uma variável se divide bem e possui um bom poder preditivo, ela é adicionada no modelo. Além disso, esse método possui uma interpretabilidade intuitiva.

2.3 Medidas de desempenho

Nesse trabalho de conclusão de curso, utilizamos a divisão do nosso banco de dados (*data-splitting*) em dois subgrupos disjuntos: treinamento e validação. O conjunto de treinamento tem a finalidade de ajustar cada um dos método utilizados (Regressão Logística e Árvore de Classificação). Por outro lado, o conjunto de validação tem a finalidade validar o desempenho dos métodos em dados que não foram utilizados para a construção dos

mesmos. Portanto, o banco de dados foi dividido em 70% para o treinamento e 30% para validação.

O *data-splitting* ajuda a evitar o viés na avaliação do desempenho do ajuste, uma vez que os dados utilizados para a construção do modelo são diferentes dos dados utilizados para a validação. Além disso, a divisão dos nossos dados em diferentes conjuntos evita o superajuste (*overfitting*) aos dados.

Para a identificação de qual metodologia realizada é a mais adequada ao nosso objetivo principal, comparamos a performance dos métodos no conjunto de validação a partir de medidas de desempenho, como: Acurácia, Precisão, Valor Preditivo Negativo, Sensibilidade (*Recall*), Especificidade e Estatística F_1, que são baseadas na Matriz de Confusão.

A matriz de confusão é uma tabela (Tabela 2.1), que possui informações para definirmos várias medidas que avaliam o desempenho de um classificador. Essa matriz é construída da seguinte forma (Izbicki e dos Santos, 2020):

Tabela 2.1: Matriz de confusão.

Valor	Valor Verdadeiro	
Predito	$Y = 0$	$Y = 1$
$Y = 0$	VN (verdadeiro negativo)	FN (falso negativo)
$Y = 1$	FP (falso positivo)	VP (verdadeiro positivo)

Com base na Tabela 2.1, é possível observar a relação entre os valores reais dos nossos dados e as classificações feitas pelos métodos de classificação utilizados. As quatro entradas de uma matriz de confusão são:

- **VN** (verdadeiro negativo): São os casos que são classificados de forma correta como pertencente à classe negativa. No nosso caso, eles correspondem ao número de indivíduos que não são diabéticos e que foram classificados como não tendo a doença.

- **FN** (falso negativo): São os casos que são classificados de forma incorreta como pertencente à classe negativa. No nosso caso, eles correspondem ao número de indivíduos que são diabéticos e que foram classificados como não tendo a doença.

- **FP** (falso positivo): São os casos que são classificados de forma incorreta como pertencente à classe positiva. No nosso caso, eles correspondem ao número de indivíduos que não são diabéticos e que foram classificados como tendo a doença.

- **VP** (verdadeiro positivo): São os casos que são classificados de forma correta como pertencente à classe positiva. No nosso caso, eles correspondem ao número de indivíduos que são diabéticos e que foram classificados como tendo a doença.

Desse modo, pela Tabela 2.1, é possível realizar o cálculo das métricas utilizadas neste trabalho de conclusão de curso. A primeira métrica utilizada é a Acurácia (A), a qual mede a proporção de indivíduos classificados corretamente em relação ao total de indivíduos. Sua expressão matemática é construída da seguinte maneira:

$$A = \frac{VN + VP}{(VN + VP + FP + FN)}.$$

No contexto deste trabalho, a acurácia corresponde a proporção de indivíduos não diabéticos que foram classificados como não diabéticos e os indivíduos diabéticos classificados como diabéticos com relação ao total de indivíduos na amostra.

A Precisão, também conhecida como Valor Preditivo Positivo (VPP), mede qual a proporção dos indivíduos classificados como positivos que foram corretamente classificados. Sendo assim, ela é expressa da seguinte maneira:

$$VPP = \frac{VP}{(VP + FP)}.$$

No nosso caso, a precisão mensura a proporção de indivíduos classificados como diabéticos que de fato são diabéticos.

O Valor Preditivo Negativo (VPN) mede a proporção dos indivíduos classificados como negativos que foram corretamente classificados. Sua expressão matemática é a seguinte:

$$VPN = \frac{VN}{(VN + FN)}.$$

No nosso contexto, o VPN mensura a proporção de indivíduos classificados como não diabéticos que de fato são não diabéticos.

Já o *Recall*, também conhecido como Sensibilidade (S), mede a proporção dos indivíduos classificados como positivos dentre os casos positivos, ou seja, mede a proporção de indivíduos classificados como diabéticos dentre aqueles que são diabéticos. Sua ex-

pressão matemática é dada por:

$$S = \frac{VP}{(VP + FN)}.$$

Por outro lado, a Especificidade (E), mede a proporção dos indivíduos classificados como negativos dentre os casos negativos, ou seja, mede a proporção de indivíduos classificados como não diabéticos dentre aqueles que não são diabéticos. Sua expressão matemática é dada por:

$$E = \frac{VN}{(VN + FP)}.$$

Por fim, a Estatística $F1$ é a média harmônica entre a Sensibilidade e a Precisão, o que fornece uma visão mais equilibrada do desempenho do modelo. Sua expressão matemática é dada por:

$$F1 = \frac{2}{\frac{1}{S} + \frac{1}{VPP}}.$$

Para esse trabalho de conclusão de curso, serão consideradas todas as métricas conjuntamente para obter uma avaliação mais completa do desempenho de cada um dos métodos em diferentes aspectos.

2.4 Desbalanceamento das classes

Em muitos estudos de classificação podemos nos deparar com o desbalanceamento de classes, que se refere à situação em que as classes que estão sendo previstas por um modelo têm distribuições muito diferentes em termos de tamanho, ou seja, existe um grande desequilíbrio entre o número de observações das diferentes classes no conjunto de dados.

Desse modo, em um exemplo de classificação binária com classes 0 e 1, em que a classe 1 tem probabilidade pequena de acontecer, a $\mathbb{P}(Y = 1|\boldsymbol{x})$ também terá o valor muito baixo, o que indica que considerar a classificação utilizando o classificador de Bayes dada por:

$$g(\boldsymbol{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\boldsymbol{x}) \geq 0.5),$$

ou seja, dado um conjunto de covariáveis $\boldsymbol{x}$, classificamos Y como 1 se a probabilidade estimada for maior do que 0.5, teremos um classificador trivial $g(\boldsymbol{x}) \equiv 0$.

Uma das formas mais utilizadas para ajustar esse problema consiste em encontrar pontos de corte K que são diferentes de 0.5 nos classificadores probabilísticos. Sendo assim, definimos a função de classificação como:

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

Como mencionado por Izbicki e dos Santos (2020), um método muito utilizado para encontrar o melhor ponto de corte K é a Curva ROC (*Receiver operating characteristic*), que é um gráfico de diagnóstico construído com o conjunto de teste, que apresenta uma curva com a variação da sensibilidade e a especificidade de forma conjunta para diferentes valores de K. No gráfico da Curva ROC existe uma linha diagonal, indo do canto inferior esquerdo para o superior direito, que indica a "curva" de um classificador ruim, que prevê a classe majoritária em todos os casos. Além disso, a partir da Curva ROC é obtida a métrica AUC (*Area Under the Curve*) que avalia a capacidade preditiva de um modelo de classificação binária, ou seja, quanto maior a AUC, melhor é a capacidade preditiva do modelo.

Existem muitas abordagens para a escolha do K usando a Curva ROC, sendo algumas delas:

- A Média Geométrica (Média-G) que é uma métrica para classificação desbalanceada que busca um equilíbrio entre a sensibilidade e a especificidade, quando é otimizada. Sua expressão matemática é dada por:

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

- A estatística de Youden (J) que é uma métrica que leva em consideração tanto a sensibilidade quanto a especificidade. Sua expressão matemática é dada por:

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

O valor de K é definido como aquele que maximiza a estatística Média-G ou a estatística de Youden.

Outra alternativa usada para a escolha do melhor ponto de corte K é A Curva de Precisão-*Recall*, que é focada apenas no desempenho de um classificador na classe positiva (classe minoritária).

A curva de Precisão e *Recall* apresenta a variação da precisão e do *recall* (sensibilidade) para diferentes K. Além disso, é apresentada uma linha horizontal com uma precisão que é a razão de casos positivos no conjunto de treinamento.

O ponto de corte que resulta no melhor equilíbrio entre a Precisão e o *Recall* é o mesmo que otimiza a mensuração F (*F-measure*) que resume a média harmônica de ambas as medidas, assim como a Estatística F1. O *F-measure* é dado por:

$$F\text{-}measure = 2 \cdot \frac{\text{Precisão} \cdot Recall}{\text{Precisão} + Recall}.$$

Desse modo, para tratarmos o desbalanceamento das classes, consideramos os métodos de escolha do ponto de corte ótimo, apresentados nessa seção.

Capítulo 3

Análise Descritiva e Exploratória

Nesse capítulo é apresentado a estruturação do banco de dados disponível e um estudo inicial dos dados para a verificação do comportamento da variável resposta e de cada uma das covariáveis a partir de análises gráficas.

3.1 Estrutura do banco de dados

Os dados que são utilizados nas análises estatísticas deste trabalho de conclusão de curso fazem parte da base de dados 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. O conjunto de dados é composto por 898 indivíduos, uma variável resposta qualitativa, 15 covariáveis qualitativas, cinco covariáveis quantitativas, totalizando 20 covariáveis escolhidas com ajuda do pesquisador.

Para auxiliar o estudo, as variáveis utilizadas são descritas abaixo como foram definidas no dicionário do projeto temático em que foi retirado o banco de dados.

A variável resposta corresponde à Diabetes, definida como:

$$Y = \begin{cases} 1, \text{ se o indivíduo é diabético;} \\ 0, \text{ se o indivíduo não é diabético.} \end{cases}$$

Para o indivíduo ser classificado como diabético, ao fazer a medição sanguínea da glicemia em jejum, ela deve ser maior do que 126mg/dL ou o indivíduo deve fazer o uso de medicação oral (hipoglicemiantes) ou de insulina injetável.

As covariáveis são apresentadas em blocos de acordo com suas características comuns,

como são apresentadas abaixo:

1. **Variáveis de identificação:**

   - **Sexo:** apresenta o sexo do indivíduo,

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

   - **Idade:** apresenta a idade contínua do indivíduo (em anos);

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

   - **Faixa Etária:** Categorização da idade do indivíduo,

$$\text{Faixa Etária} = \begin{cases} 0, \text{ se Adolescente (0 à 19 anos);} \\ 1, \text{ se Adulto (20 à 59 anos);} \\ 2, \text{ se Idoso (60 anos ou mais).} \end{cases}$$

     A variável Faixa Etária não é utilizada na aplicação do estudo e é apenas utilizada para melhor visualização da idade do indivíduo em categorias na análise descritiva e exploratória dos dados.

   - **Índice de massa corporal (IMC):** apresenta a relação entre o peso e a altura de uma pessoa. Seu cálculo é feito de acordo com a faixa etária do indivíduo com base na definição do pesquisador no dicionário do projeto. Baseado no resultado do IMC, o indivíduo é alocado em uma categoria. Cada uma das categorias é apresentada a seguir:

$$\text{IMC} = \begin{cases} 0, \text{ se o indivíduo está abaixo do peso;} \\ 1, \text{ se o indivíduo está com o peso ideal (eutrofia);} \\ 2, \text{ se o indivíduo está acima do peso (sobrepeso);} \\ 3, \text{ se o indivíduo é obeso.} \end{cases}$$

3. **Doenças crônicas:**

- **Hipertensão Arterial:** apresenta a condição de um indivíduo que possui a pressão do sangue nas artérias cronicamente elevada. Sua classificação é feita de acordo com a faixa etária do indivíduo:

$$\text{Hipertensão Arterial} = \begin{cases} 1, \text{ se o indivíduo possui hipertensão arterial;} \\ 0, \text{ se o indivíduo não possui hipertensão arterial.} \end{cases}$$

- **Colesterol não-HDL:** refere-se à diferença entre o colesterol total e o colesterol HDL (lipoproteína de alta densidade). O colesterol não-HDL inclui o colesterol LDL (lipoproteína de baixa densidade), conhecido como "colesterol ruim", pois ele está associado ao risco de doenças cardiovasculares. A classificação do colesterol não-HDL é feita de acordo com a idade do indivíduo:

$$\text{Colesterol não-HDL} = \begin{cases} 1, \text{ inadequado;} \\ 0, \text{ adequado.} \end{cases}$$

- **Colesterol HDL (lipoproteína de alta densidade):** se refere a uma fração do colesterol presente no sangue. É conhecido como "colesterol bom", pois é importante para a remoção de excesso de colesterol nas artérias e células. É necessário ter níveis adequados de colesterol HDL para um menor risco de doenças cardiovasculares. A sua classificação é feita de acordo com a idade e o sexo do indivíduo:

$$\text{Colesterol HDL} = \begin{cases} 1, \text{ inadequado;} \\ 0, \text{ adequado.} \end{cases}$$

- **Dislipidemia:** se refere a uma elevação nos níveis de gorduras (lipídios) no sangue, incluindo o colesterol e os triglicerídeos. Sua classificação é dada por:

$$\text{Dislipidemia} = \begin{cases} 1, \text{ se o indivíduo possui dislipidemia;} \\ 0, \text{ se o indivíduo não possui dislipidemia.} \end{cases}$$

- **Doença mental ou problema emocional:** indica se o indivíduo tem algum tipo de doença mental ou problema emocional como ansiedade, depressão,

TOC, esquizofrenia:

$$\text{Doença mental ou problema emocional} = \begin{cases} 1, \text{ se o indivíduo não possui;} \\ 2, \text{ se o indivíduo possui.} \end{cases}$$

4. **Comportamentos relacionados à saúde:**

- **Tabagismo:** indica o hábito de consumo de tabaco:

$$\text{Tabagismo} = \begin{cases} 1, \text{ se o indivíduo é ex-fumante ou fumante;} \\ 0, \text{ se o indivíduo nunca fumou.} \end{cases}$$

- **Consumo de bebida alcoólica:** apresenta se o indivíduo consome bebida alcoólica:

$$\text{Consumo de bebida alcoólica} = \begin{cases} 0, \text{ se o indivíduo nunca bebeu;} \\ 1, \text{ se o indivíduo parou de beber;} \\ 2, \text{ se o indivíduo bebe atualmente.} \end{cases}$$

5. **Prática de atividade física:**

- **Atividade física no lazer:** indica se o indivíduo cumpre a recomendação (em minutos) de prática de atividade física no lazer por semana. A recomendação para adolescentes é de 420 minutos por semana, para adultos e idosos é de 150 minutos por semana.

$$\text{Atividade física no lazer} = \begin{cases} 1, \text{ se o indivíduo cumpre a recomendação;} \\ 0, \text{ se o indivíduo não cumpre a recomendação.} \end{cases}$$

- **Atividade física doméstica:** indica se o indivíduo cumpre a recomendação (em minutos) de prática de atividade física doméstica por semana. A recomendação para adolescentes é de 420 minutos por semana, para adultos e

idosos é de 150 minutos por semana.

$$\text{Atividade física doméstica} = \begin{cases} 1, & \text{se o indivíduo cumpre a recomendação;} \\ 0, & \text{se o indivíduo não cumpre a recomendação.} \end{cases}$$

- **Atividade física no transporte:** indica se o indivíduo cumpre a recomendação (em minutos) de prática de atividade física no transporte por semana. A recomendação para adolescentes é de 420 minutos por semana, para adultos e idosos é de 150 minutos por semana.

$$\text{Atividade física no transporte} = \begin{cases} 1, & \text{se o indivíduo cumpre a recomendação;} \\ 0, & \text{se o indivíduo não cumpre a recomendação.} \end{cases}$$

- **Atividade física no trabalho:** indica se o indivíduo cumpre a recomendação (em minutos) de prática de atividade física no trabalho por semana. A recomendação para adolescentes é de 420 minutos por semana, para adultos e idosos é de 150 minutos por semana.

$$\text{Atividade física no trabalho} = \begin{cases} 1, & \text{se o indivíduo cumpre a recomendação;} \\ 0, & \text{se o indivíduo não cumpre a recomendação.} \end{cases}$$

6. **Entretenimento Tecnológico:**

- **Tempo assistindo TV durante um dia da semana:** apresenta quanto tempo no total (em horas) o indivíduo gasta assistindo TV durante um dia da semana.
- **Tempo assistindo TV durante um dia do final de semana:** apresenta quanto tempo no total (em horas) o indivíduo gasta assistindo TV durante um dia de final de semana.
- **Tempo no computador durante um dia da semana:** apresenta quanto tempo no total (em horas) o indivíduo gasta no computador durante um dia da semana.
- **Tempo no computador durante um dia do final de semana:** apresenta quanto tempo no total (em horas) o indivíduo gasta no computador durante

um dia de final de semana.

3.2 Análise descritiva e exploratória dos dados

A Análise descritiva e exploratória do conjunto de dados é essencial para compreender o comportamento de cada uma das variáveis em estudo. Além disso, a partir dessa análise podemos identificar padrões, pontos *outliers* e resumir a informação de cada variável em estudo.

Primeiramente, foi realizada a identificação da existência de valores faltantes e não respostas na variável resposta e em cada uma das covariáveis, apresentada na Tabela 3.1.

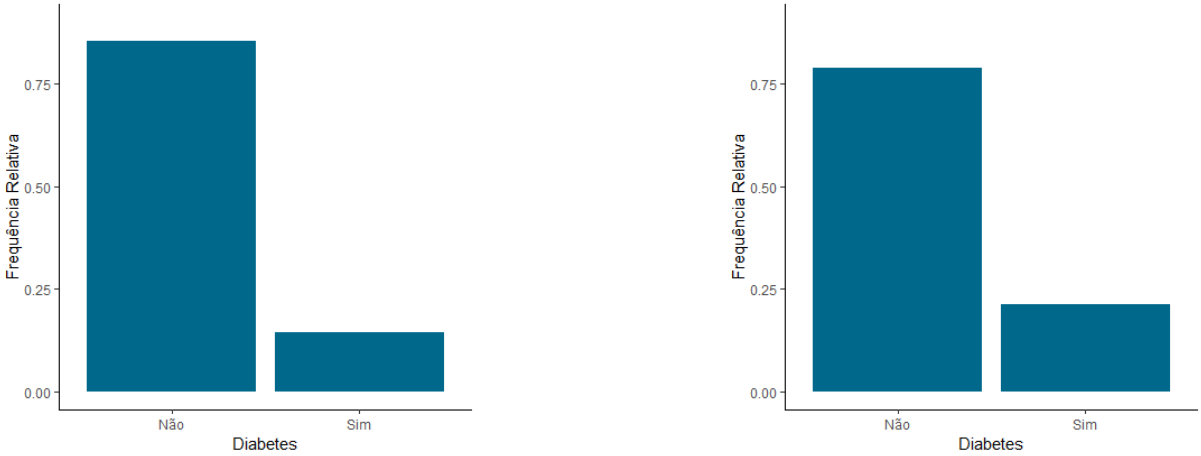
Tabela 3.1: Quantidade de valores faltantes em cada uma das variáveis.

Variável	**Quantidade de valores faltantes**
Diabetes	0
Sexo	0
Idade	0
Faixa Etária	0
IMC	12
Hipertensão Arterial	10
Colesterol não-HDL	19
Colesterol HDL	19
Dislipidemia	14
Doença Mental ou Problema Emocional	10
Tabagismo	6
Consumo de bebida alcoólica	8
Atividade física no lazer	5
Atividade física doméstica	13
Atividade física no transporte	5
Atividade física no trabalho	23
Tempo assistindo TV (um dia da semana)	6
Tempo assistindo TV (um dia do final de semana)	9
Tempo no computador (um dia da semana)	7
Tempo no computador (um dia do final de semana)	6

Averiguando os resultados obtidos na Tabela 3.1, notamos que não há valores faltantes na variável resposta (Diabetes). Ademais, além da variável de atividade física no trabalho, as variáveis que possuem um maior número de valores faltantes são as que pertencem ao bloco de doenças crônicas, o que é esperado uma vez que para a obtenção dessas informações, fez-se necessária a coleta de sangue dos indivíduos. Apesar de existirem dados faltantes em algumas variáveis, é uma quantidade baixa comparada ao número total de observações (898 indivíduos). Desse modo, optamos por retirar os indivíduos que

possuem valores faltantes para continuarmos o estudo. Portanto, nosso banco de dados passa a possuir 829 indivíduos.

Em seguida, para maior entendimento do comportamento da variável resposta, foi realizada a construção de um gráfico de barras (Figura 3.1a) com a frequência relativa dos valores que a variável pode assumir. Além disso, fisiologicamente, é esperado ter prevalência de diabetes para indivíduos com idades mais altas. Portanto, realizamos a construção do gráfico de barras removendo os indivíduos que compõe a faixa etária de adolescentes (Figura 3.1b), para fins de comparação da variável diabetes. Desse modo, removendo os adolescentes do nosso conjunto de dados, passamos a possuir 608 observações.



(a) Considerando todos os indivíduos. (b) Desconsiderando os adolescentes.

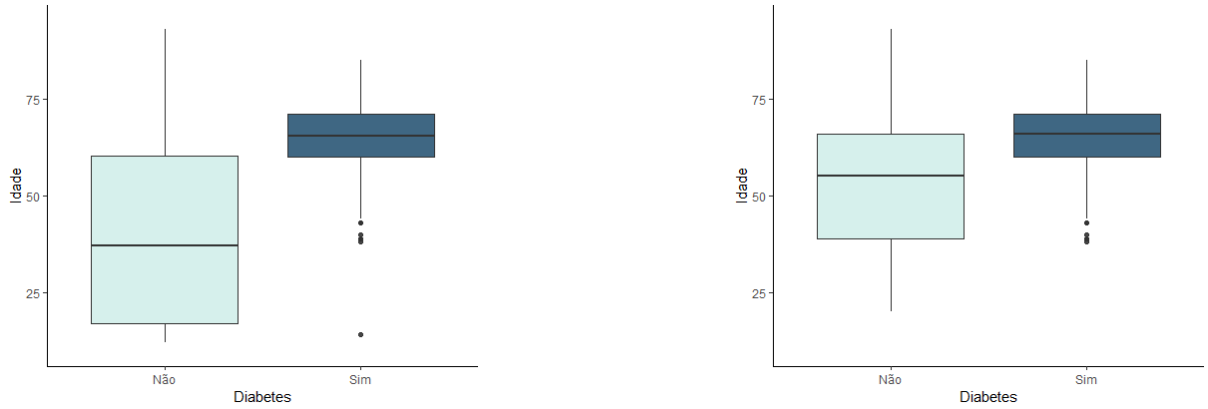
Figura 3.1: Frequência relativa da variável resposta Diabetes.

Observamos pela Figura 3.1, que em ambos os gráficos nosso banco de dados é composto em sua maioria por indivíduos que não são diabéticos(3.1a e 3.1b), com uma frequência relativa de 0,86 e 0,80, respectivamente, o que indica que o conjunto de dados é desbalanceado. Além disso, se desconsiderarmos os adolescentes da nossa análise, como visto no Gráfico 3.1b, passamos de uma percentual de $14,5\%$ indivíduos com diabetes para $21,2\%$, ou seja, não temos uma grande mudança no comportamento da variável resposta analisada individualmente quando é desconsiderada a faixa etária de idades mais baixas do banco de dados.

Para um melhor entendimento das variáveis em estudo, realizamos uma análise gráfica que ilustra a influência de cada uma das covariáveis na variável resposta. Para essa finalidade, para as variáveis explicativas categóricas, construímos gráficos de barras com a frequência relativa da variável Diabetes, considerando cada categoria da covariável. Já para as variáveis explicativas quantitativas, construímos gráficos boxplot ou histograma

da variável Diabetes em relação a covariável. Além disso, para as análises gráficas foi considerado apenas os dados observados, sendo assim os dados faltantes de cada variável não foram incluídos nessa análise.

Notamos que algumas variáveis possuem seu comportamento influenciado pela presença ou não de indivíduos adolescentes no banco de dados, sendo elas: Idade, Hipertensão Arterial, Tabagismo e Consumo de bebida alcoólica. Desse modo, construímos uma análise gráfica para esses casos considerando os adolescente e outra desconsiderando-os. Primeiramente, construímos os boxplots para a variável diabetes em relação à idade, representados na Figura 3.2.

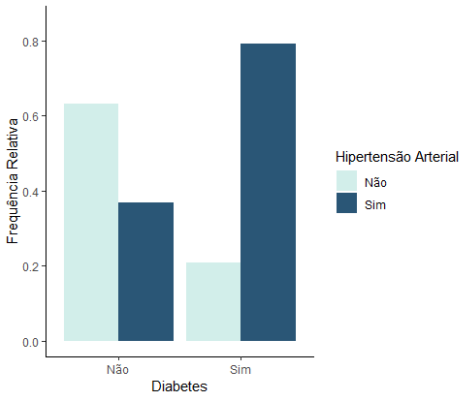


(a) Considerando todos os indivíduos. (b) Desconsiderando os adolescentes.

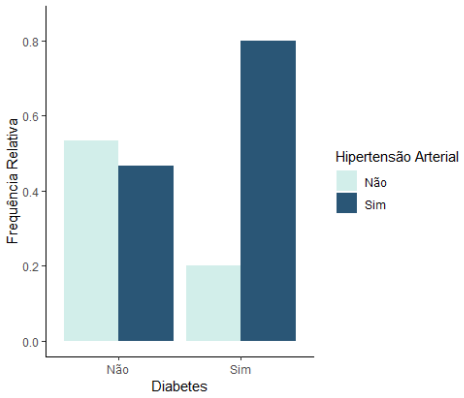
Figura 3.2: Boxplot da variável Diabetes em relação a covariável idade.

Assim como observado na Figura 3.2, quando consideramos todos os indivíduos (Figura 3.2a) não há intersecção entre as caixas de indivíduos diabéticos e não diabéticos, o que indica que a diabetes pode ser influenciada pela idade do indivíduo, o que é esperado, uma vez que indivíduos diabéticos possuem idade concentrada em valores maiores com apenas alguns *outliers* em idades mais baixas e indivíduos não diabéticos possuem idades mais baixas. Quando comparamos o gráfico em que são desconsiderados os adolescentes (Figura 3.2b) com o gráfico considerando todos os indivíduos (Figura 3.2a), notamos que não há diferenças para indivíduos diabéticos, apenas a perda de um *outlier* inferior quando desconsideramos os adolescente e para indivíduos que não são diabéticos, notamos que a caixa do boxplot fica mais concentrada entre idades de 42 à 68 anos e a mediana muda de 43 anos para 60 anos.

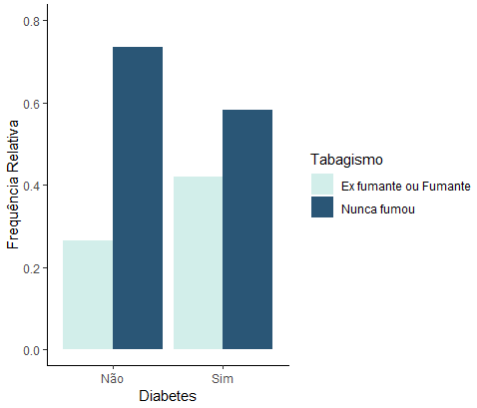
Posteriormente, construímos gráficos de barras de frequência relativa da variável Diabetes em relação as categorias das variáveis explicativas categóricas, apresentados na Figura 3.3.



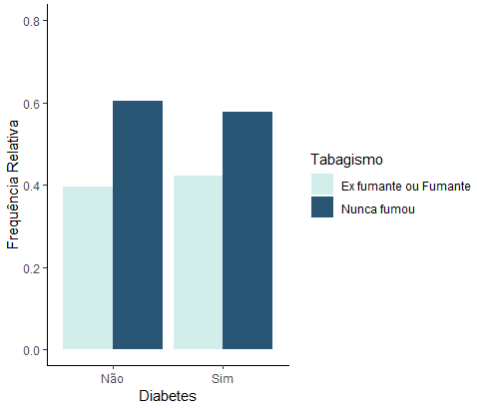
(a) Hipertensão Arterial considerando todos os indivíduos.



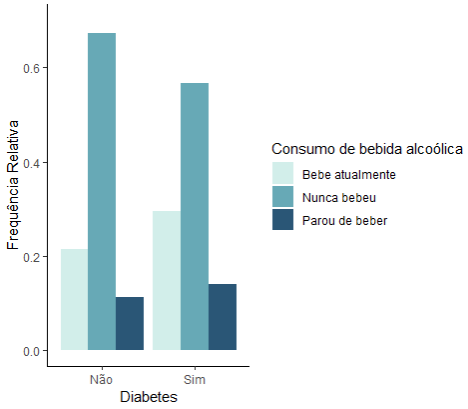
(b) Hipertensão Arterial desconsiderando os adolescentes.



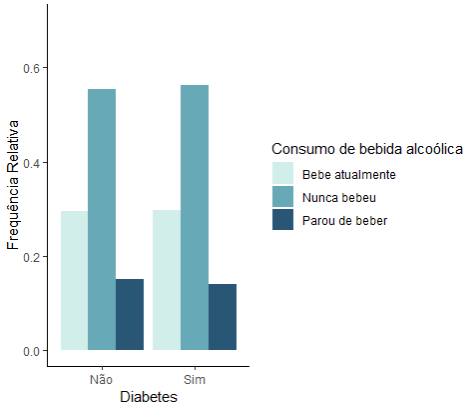
(c) Tabagismo considerando todos os indivíduos.



(d) Tabagismo desconsiderando os adolescentes.



(e) Consumo de bebida alcoólica considerando todos os indivíduos.



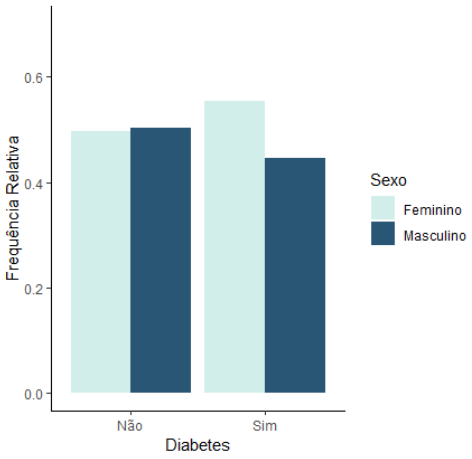
(f) Consumo de bebida alcoólica desconsiderando os adolescentes.

Figura 3.3: Gráficos de barras da proporção de diabetes em relação as categorias das variáveis que possuem o comportamento influenciado pelos adolescentes no banco de dados.

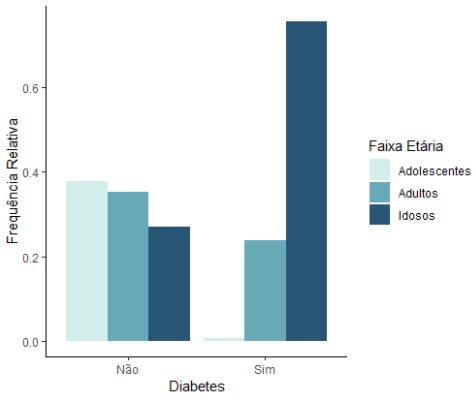
É interessante observar na Figura 3.3, que quando os adolescentes são desconsiderados do banco de dados, a Hipertensão Arterial para indivíduos que não são diabéticos apresenta maior proporção quando comparamos com a proporção considerando todos os indivíduos. Notamos também que tanto quando consideramos todo o banco de dados

(Figura 3.3a) e quando desconsideramos os adolescentes (Figura 3.3b), a maior parte dos indivíduos que são diabéticos também possuem hipertensão arterial. Por fim, o Tabagismo e o Consumo de bebida alcoólica, quando desconsideramos os adolescentes, são proporcionais e parecidos para indivíduos diabéticos e não diabéticos, o que indica que essas variáveis podem não trazer muita informação para a predição de diabetes.

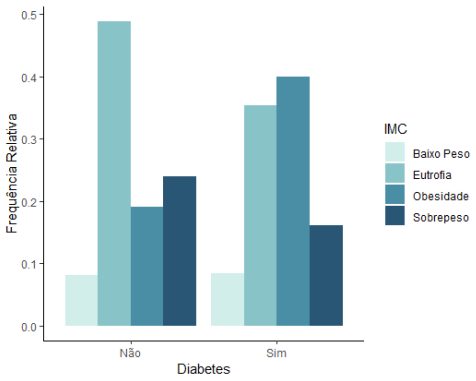
Após essa análise, prosseguimos em estudar o comportamento da variável resposta em relação as demais covariáveis que não possuem influência pela presença ou não de adolescentes no conjunto de dados. Desse modo, construímos gráficos de barras de frequência relativa da variável Diabetes em relação as categorias de cada uma dessas variáveis, considerando todos os indivíduos do conjunto de dados. Na Figura 3.4 apresentamos os gráficos da variável resposta em relação as variáveis dos blocos 1 e 2, sendo eles Variáveis de identificação e Variáveis criadas a partir do questionário.



(a) Sexo.



(b) Faixa Etária.



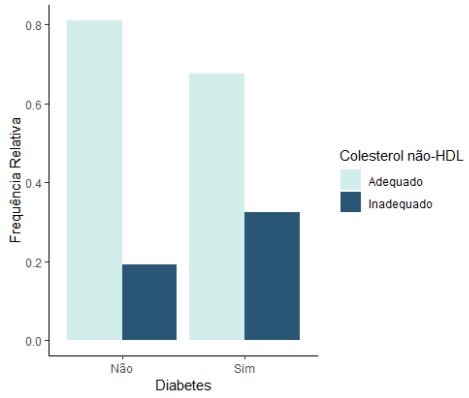
(c) IMC.

Figura 3.4: Gráfico de barras da proporção de indivíduos diabéticos ou não diabéticos em cada uma das categorias das variáveis pertencentes aos blocos 1 e 2.

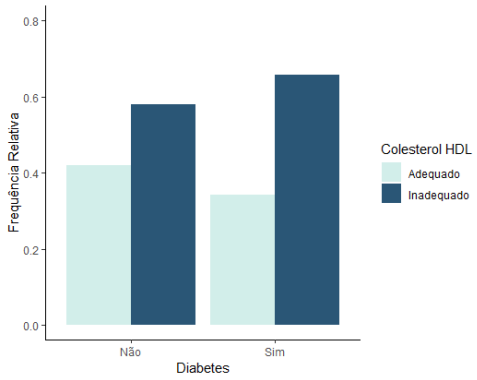
Notamos na Figura 3.4 que a diabetes é mais frequente para indivíduos do sexo feminino. Ademais, é observado que indivíduos não diabéticos são, aproximadamente, propor-

cionais em relação ao sexo. Em relação a faixa etária (Figura 3.4b), percebemos que essa variável pode ser significativa para predizer a diabetes, pois através do gráfico é observado que a diabetes é predominante em idosos, seguido por adultos. Além disso, grande parte de indivíduos não diabéticos são adolescentes ou adultos. Por fim, como visto na Figura 3.4c, indivíduos não diabéticos possuem, em sua maioria, o peso ideal (eutrofia), enquanto 40% dos indivíduos diabéticos são obesos seguidos por 35% que possuem o peso ideal (eutrofia).

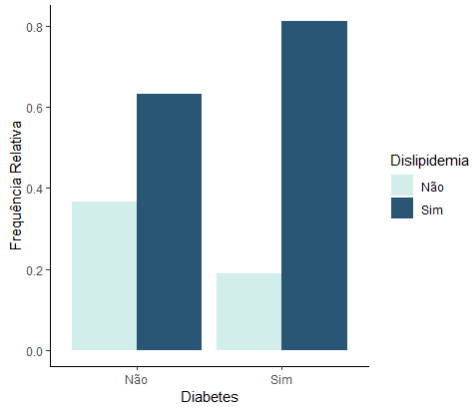
Logo após, na Figura 3.5 apresentamos os gráficos da variável resposta em relação as variáveis do bloco 3 (Doenças crônicas).



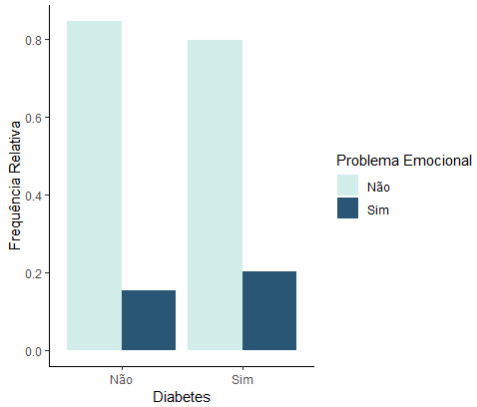
(a) Colesterol não-HDL.



(b) Colesterol HDL.



(c) Dislipidemia.



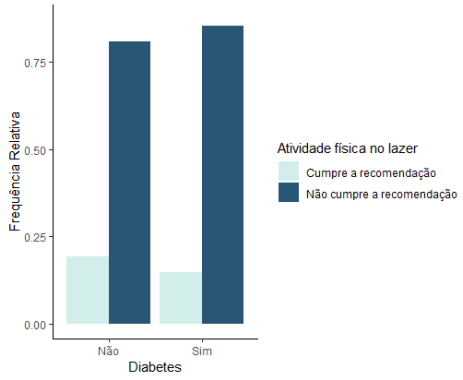
(d) Problema Emocional.

Figura 3.5: Gráfico de barras da proporção de indivíduos diabéticos ou não diabéticos em cada uma das categorias das variáveis pertencentes ao bloco 3.

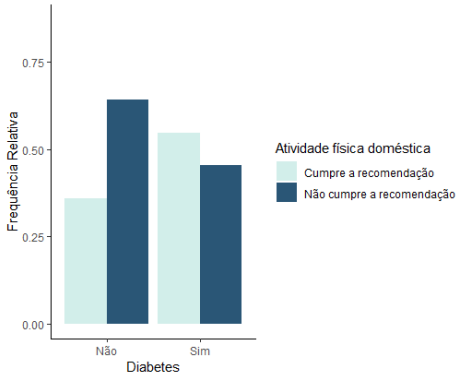
Como é observado na Figura 3.5, tanto para indivíduos não diabéticos quanto para indivíduos diabéticos, o Colesterol não-HDL na categoria adequado possui maior proporção, o que indica que possivelmente essa variável não seja muito significativa para predizer a diabetes. Esse comportamento também é visto nas variáveis Colesterol HDL, Dislipidemia e Problema Emocional, as quais possuem maior proporção de inadequado, sim e não,

respectivamente, para ambas categorias da variável resposta.

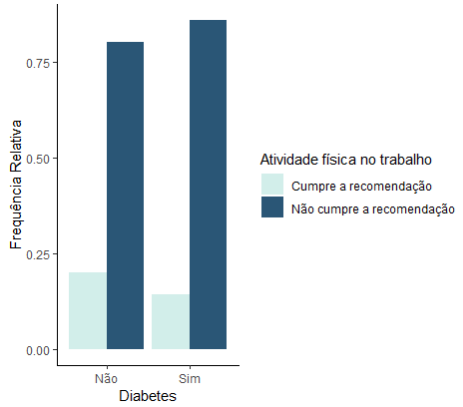
Na sequência, foi construído a representação gráfica da variável resposta em relação as categorias das variáveis do bloco 5 (Prática de atividade física), Figura 3.6.



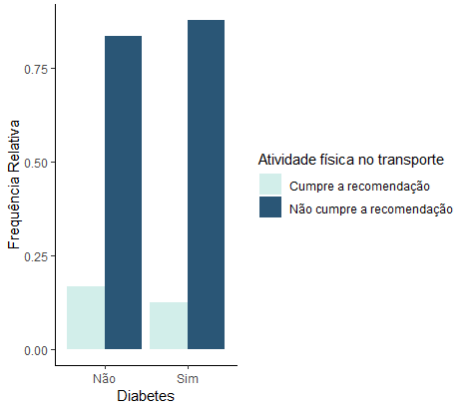
(a) Atividade física no lazer.



(b) Atividade física doméstica.



(c) Atividade física no trabalho.



(d) Atividade física no transporte.

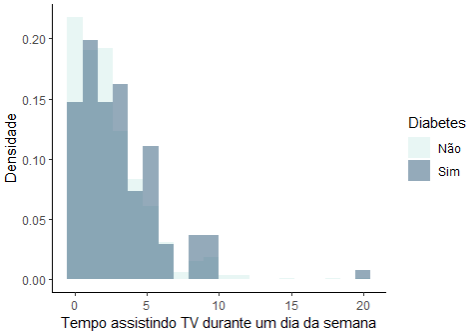
Figura 3.6: Gráfico de barras da proporção de indivíduos diabéticos ou não diabéticos em cada uma das categorias das variáveis pertencentes ao bloco 5.

Por meio da Figura 3.6, notamos que para ambas as categorias da variável resposta, a maior parte dos indivíduos não cumpre a recomendação de minutos semanais de prática de atividade física no lazer, transporte e trabalho, como é ilustrado nas Figuras 3.6a, 3.6c e 3.6d, respectivamente. Esse comportamento evidencia que essas variáveis podem não possuir muita influência para predizer a diabetes. Por fim, na Figura 3.6b, observamos que 64% dos indivíduos não diabéticos, não cumprem a recomendação de atividade física doméstica, enquanto 55% dos indivíduos diabéticos cumprem a recomendação.

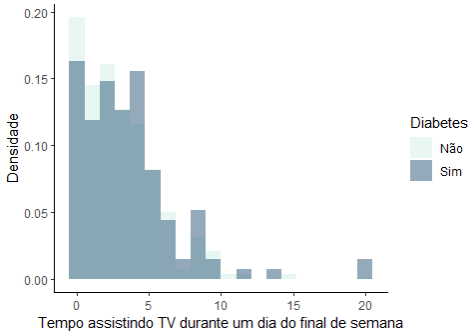
Em seguida, construímos gráficos para analisarmos o comportamento da variável resposta em relação as covariáveis pertencentes ao bloco 6 (Entretenimento tecnológico), 3.7.

Ao observar a Figura 3.7, analisamos, primeiramente a Figura 3.7a que a distribuição dos dados é muito parecida para indivíduos diabéticos e não diabéticos, em que a maior

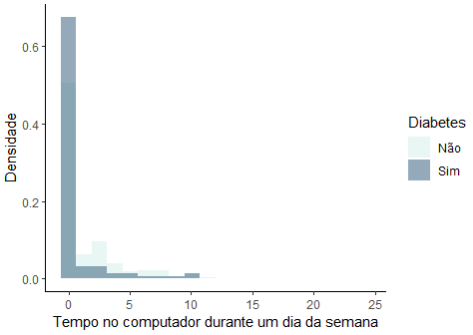
parte dos indivíduos assistem TV durante um dia da semana entre 0 e 5 horas e há a presença de poucos *outliers*. Desse modo, não temos evidências de que o tempo gasto assistindo TV durante um dia da semana tem forte influência na variável resposta. Na Figura 3.7b, notamos um comportamento parecido, em que há valores *outliers* para ambas as categorias da variável Diabetes, e também a maior parte dos indivíduos assistem TV em um dia do final de semana entre 0 e 5 horas. Pela semelhança do comportamento para indivíduos diabéticos e não diabéticos, temos evidências de que não existe forte influência do tempo assistindo TV durante um dia do final de semana na variável resposta. Por fim, analisando as Figuras 3.7c e 3.7d, notamos que elas são semelhantes entre si, desse modo, o tempo no computador durante um dia da semana e durante um dia do final de semana é parecido para os indivíduos. Além disso, podemos destacar que a maior parte dos indivíduos diabéticos não usa o computador tanto em dia de semana quanto no final de semana (com excessão dos *outliers*), pois grande parte dos indivíduos estão concentrados em zero em ambos os casos.



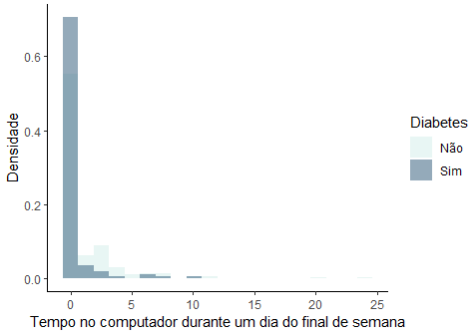
(a) Tempo assistindo TV durante um dia da semana (horas).



(b) Tempo assistindo TV durante um dia do final de semana (horas).



(c) Tempo no computador durante um dia da semana (horas).



(d) Tempo no computador durante um dia do final de semana (horas).

Figura 3.7: Histogramas da variável Diabetes em relação as covariáveis pertencentes ao bloco 6.

Em última análise, construímos o gráfico da matriz de correlação entre nossas variáveis

quantitativas com a finalidade de detectar a correlação entre elas. Para isso, denotamos as variáveis contínuas da seguinte maneira:

- TV (semana): tempo assistindo TV (em horas) em um dia da semana;
- TV (final de semana): tempo assistindo TV (em horas) em um dia do final de semana;
- Computador (semana): tempo no computador (em horas) em um dia da semana
- Computador (final de semana): tempo no computador (em horas) em um dia do final de semana;
- Idade: idade do indivíduo (em anos).

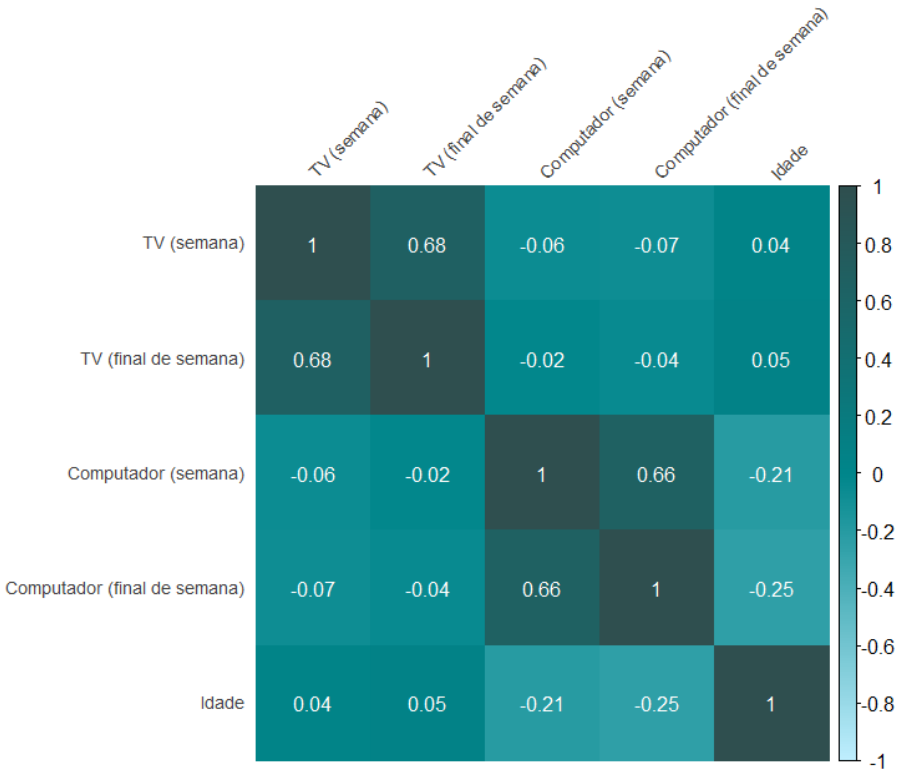


Figura 3.8: Gráfico da matriz de correlação entre as covariáveis contínuas.

Na Figura 3.8 observamos que a correlação entre o tempo assistindo TV em um dia da semana e o tempo assistindo TV em um dia do final de semana é de 0.68, o que indica que a correlação entre essas duas variáveis é moderada, sendo assim optamos em manter essas duas covariáveis. Além disso, a correlação entre o tempo no computador em um dia da semana e o tempo no computador em um dia do final de semana é de 0.66, que

também é considerada uma correlação moderada. Desse modo, como não possuímos uma correlação alta, mantemos essas variáveis. As demais correlações são consideradas baixas e, portanto, mantemos todas as covariáveis do nosso banco de dados.

Capítulo 4

Aplicação

Nesse capítulo é apresentada a divisão do banco de dados em dois subgrupos disjuntos, treinamento e validação, em que o conjunto de treinamento é utilizado para o ajuste dos métodos e o conjunto de validação é utilizado para avaliar o desempenho dos mesmos. Além disso, são apresentados os resultados das aplicações dos métodos de Regressão Logística e Árvore de Decisão, e exibido suas medidas de desempenho.

4.1 Divisão dos dados

Com base na análise descritiva realizada no Capítulo 3 não foi observado diferença na análise quando desconsideramos os adolescentes do estudo. Desse modo, consideramos todos os indivíduos para realizar a análise. Foi utilizado uma divisão aleatória dos dados em treinamento e validação. Portanto, o banco de dados foi dividido em 70% para o treinamento (581 indivíduos) e 30% para a validação (248 indivíduos).

Na Tabela 4.1 é apresentada a proporção das categorias da variável resposta no conjunto de treinamento e no conjunto de validação. Desse modo, observamos as proporções de indivíduos diabéticos e não diabéticos foi mantida igual no treinamento e na validação para melhor confiança nas generalizações dos modelos e para a igualdade de representabilidade dos dados.

4.2 Regressão Logística

Como discutido no Capítulo 2, propomos utilizar a Regressão Logística para predizer nossa variável resposta Y que assume 1, se o indivíduo for diabético, ou 0, caso contrário.

Tabela 4.1: Proporção das categorias da variável resposta nos conjuntos de treinamento e validação.

Diabetes	**Proporção**	
	Conjunto de Treinamento	Conjunto de Validação
Sim	0,14	0,14
Não	0,86	0,86

Com essa finalidade, consideramos implementar a Regressão Logística utilizando todas as covariáveis disponíveis no banco de dados e também implementamos utilizando apenas covariáveis selecionadas pelo método *Stepwise*. Além disso, implementamos ambos os ajustes utilizando a linguagem de programação R (R Development Core Team, 2020).

4.2.1 Regressão Logística utilizando todas covariáveis

Primeiramente, da maneira em que é demonstrado na Seção 2.1, implementamos a Regressão Logística utilizando todas as covariáveis disponibilizadas no banco de dados.

Após o ajuste do modelo, realizamos as predições das observações do conjunto de validação e obtivemos as estimativas das probabilidades, $\hat{\mathbb{P}}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta})$. Desse modo, para construir um classificador $g(\boldsymbol{x})$, inicialmente, consideramos um ponto de corte K igual à $0,5$, porém, como já foi mencionado, temos desbalanceamento entre as classes. Portanto, consideramos também as abordagens para a escolha do K descritas na Seção 2.4.

As duas abordagens iniciais para a escolha do K são baseadas na Curva ROC. Sendo assim, apresentamos a Curva ROC na Figura 4.1.

Com base na Figura 4.1, consideramos a métrica AUC (Área Sob a Curva) que avalia o desempenho de um modelo de classificação. Portanto, como possuímos um AUC= 0,8, consideramos que o modelo possui um bom desempenho em discriminar as classes da variável resposta. Além disso podemos observar que quando aumentamos a sensibilidade, diminuímos a especificidade. Sendo assim, como uma forma de balancear a sensibilidade e a especificidade, utilizamos as métricas Média-G e J e obtivemos o mesmo ponto de corte igual a 0,104, que otimizou ambas as métricas.

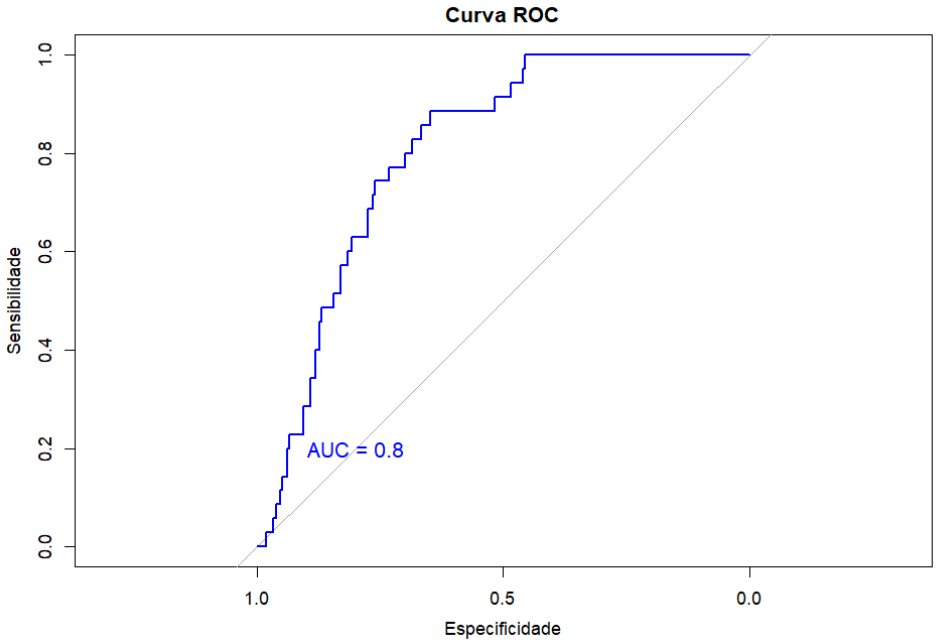


Figura 4.1: Curva ROC das predições do ajuste de Regressão Logística com todas covariáveis.

Outra abordagem para a escolha do ponto de corte K é a *F-measure* baseada na Curva de Precisão-*Recall*. Desta maneira, apresentamos a Curva de Precisão-*Recall* na Figura 4.2:

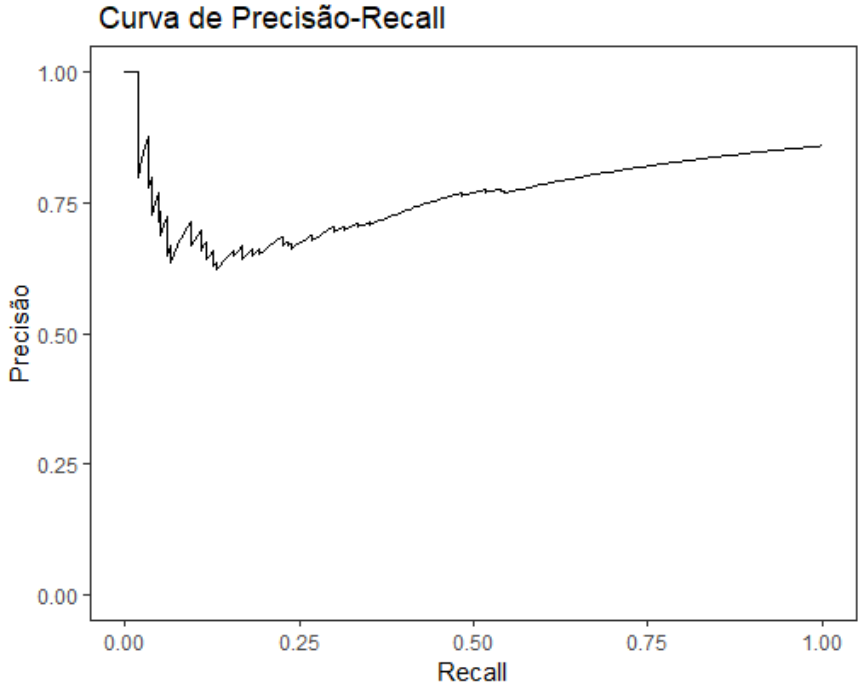


Figura 4.2: Curva de Precisão-*Recall* das predições do ajuste de Regressão Logística com todas covariáveis.

Através da Figura 4.2, notamos que apesar de termos uma maior precisão quando o

recall é zerado, quando chegamos em um *recall* próximo de $0,25$, a precisão começa a aumentar novamente. Desse modo, a *F-measure* tem a finalidade de balancear a precisão e o *Recall* e encontramos um valor ótimo de $K = 0,001$.

Desse modo, apresentamos na Figura 4.3 as três métricas utilizadas para obter os pontos de corte.

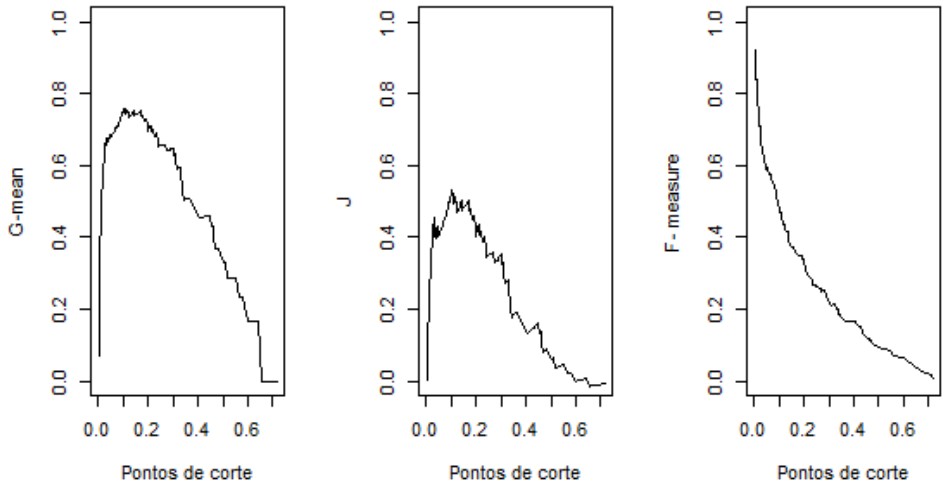


Figura 4.3: Pontos de corte para as métricas Média-G, J e *F-measure*, respectivamente.

Com base na Figura 4.3, podemos observar o ponto que maximiza cada uma das métricas, sendo assim, como já citado, função de *F-measure* está sendo maximizada no lugar diferente das funções Média-G e J, como podemos visualizar através do formato das curvas.

Após obter os classificadores com os três pontos de corte diferentes, construímos uma matriz de confusão com a finalidade de comparar as proporções de acertos em todos os cenários, descrita na Tabela 4.2.

Tabela 4.2: Matriz de confusão para classificadores obtidos pela Regressão Logística.

Valor Predito	**Valor Verdadeiro**					
	Ponto de corte = 0,5		Ponto de corte = 0,104		Ponto de corte = 0,001	
	Y = 0	Y = 1	Y = 0	Y = 1	Y = 0	Y = 1
Y = 0	0,81	0,13	0,56	0,02	0	0
Y = 1	0,04	0,02	0,3	0,12	0,86	0,14

Tabela 4.3: Medidas de Desempenho para os diferentes pontos de corte.

Medidas de Desempenho	**Ponto de Corte**		
	0,5	0,104	0,001
Acurácia	0,830	0,681	0,145
Sensibilidade (*Recall*)	0,114	0,885	1,000
Especificidade	0,948	0,648	0,004
Precisão	0,267	0,292	0,141
Valor Preditivo Negativo	0,867	0,972	1,000
Estatística F1	0,160	0,439	0,248

Com ajuda do pesquisador foi identificado que o objetivo principal é classificar de forma correta indivíduos que possuem diabetes, ou seja, queremos acertar mais os verdadeiros positivos, porém, mesmo assim precisamos de um estimador que tenha uma especificidade relativamente alta para também identificarmos corretamente os verdadeiros negativos.

Através da Tabela 4.2, notamos que para o ponto de corte usual igual à $0,5$, temos uma proporção de $0,83$ indivíduos classificados de forma correta, sendo que 81% dos indivíduos não diabéticos são classificados como não diabéticos corretamente e apenas 2% dos indivíduos diabéticos são classificados como diabéticos. Além disso, 4% dos indivíduos não diabéticos são classificados como diabéticos e 13% dos indivíduos diabéticos são classificados como não diabéticos. Para o ponto de corte $0,104$, o qual foi obtido pelas métricas de G-Média e J, obtivemos uma proporção total de acertos de $0,68$, sendo que 56% dos indivíduos não diabéticos são classificados como não diabéticos e 3% dos indivíduos não diabéticos são classificados como diabéticos. Ademais, 12% dos indivíduos diabéticos são classificados como diabéticos de forma correta e 2% dos indivíduos diabéticos são classificados como não diabéticos. Apesar da proporção total de acertos ser menor, quando comparamos com a proporção de acertos para o ponto de corte de $0,5$, temos uma sensibilidade maior, como observado na Tabela 4.3, ou seja, classificamos $88,5$% de forma correta indivíduos diabéticos. Por fim, considerando o ponto de corte $0,001$, a proporção total de acertos é de $0,14$, sendo que apenas os indivíduos diabéticos são classificados de forma correta (14%), já os indivíduos não diabéticos são classificados de forma incorreta.

Com base na análise feita e nas medidas de desempenho na Tabela 4.3, observamos que para o ponto de corte igual à $0,5$, temos valores da acurácia, da especificidade e do valor preditivo negativo altos, o que já era esperado pois a classificação da classe majoritária possui poucos erros, porém, a sensibilidade, a precisão e a Estatística F1 são muito baixas. Quando observamos as medidas de desempenho para o ponto de corte $0,104$,

notamos que seus valores são mais proporcionais, ou seja, possuímos uma sensibilidade razoavelmente alta e também uma sensibilidade alta, desse modo, acertamos mais os indivíduos que são diabéticos, que é nosso principal objetivo, mas também temos uma taxa de acertos razoavelmente boa para indivíduos não diabéticos. Em última análise, para o ponto de corte $0,001$, temos um valor de sensibilidade máximo, pois todos os indivíduos diabéticos foram classificados de forma correta, mas também erramos na classificação de todos os indivíduos não diabéticos, desse modo as medidas de desempenho como acurácia, especificidade, precisão e estatística F1 são muito baixas, por isso a escolha desse ponto de corte não é ideal para o estudo.

4.2.2 Regressão Logística utilizando covariáveis selecionadas pelo método Stepwise

Por conseguinte, implementamos a Regressão Logística utilizando as covariáveis selecionadas pelo método Stepwise com técnica Bidirecional. Desse modo, as covariáveis selecionadas foram:

- Idade;
- IMC;
- Hipertensão Arterial;
- Colesterol não-HDL;
- Dislipidemia.

Depois de ajustarmos o modelo apenas com as covariáveis selecionadas, calculamos as estimativas utilizando as observações do conjunto de validação, resultando nas probabilidades $\hat{\mathbb{P}}(Y = 1|\boldsymbol{x}, \boldsymbol{\beta})$. Assim, ao construir o classificador $g(\boldsymbol{x})$, inicialmente também adotamos um ponto de corte K estabelecido em $0,5$. Além disso, devido ao desbalanceamento entre as classes, também exploramos abordagens adicionais para a escolha de K, conforme discutido na Seção 2.4.

Desse modo, ajustamos a Curva ROC, em que são baseadas as métricas Média-G e J, na Figura 4.4.

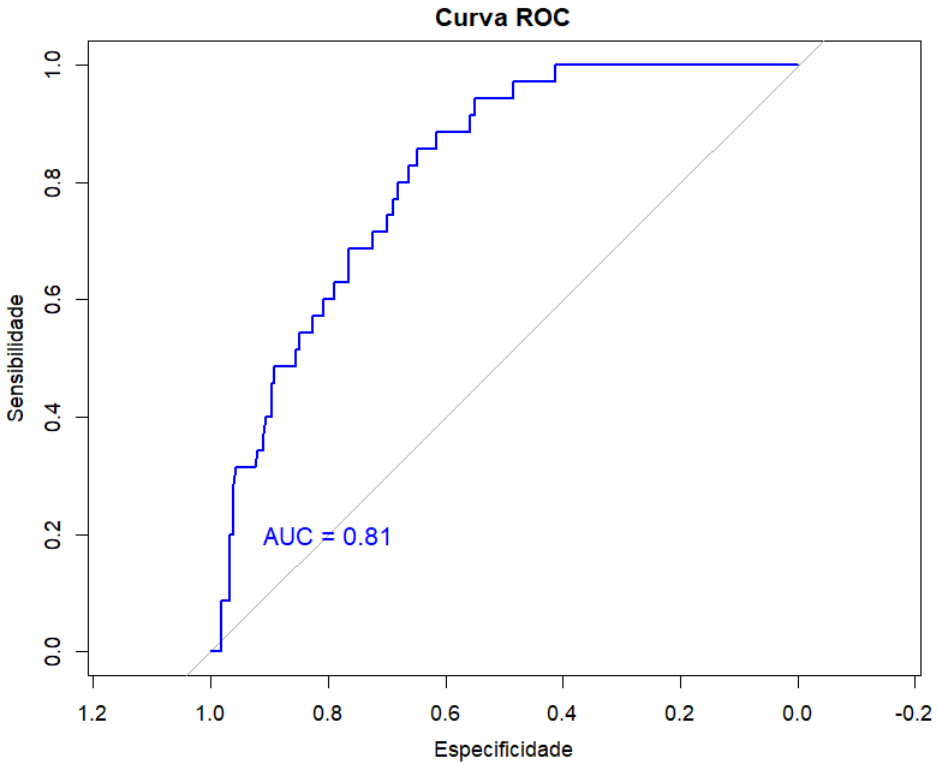


Figura 4.4: Curva ROC das predições do ajuste de Regressão Logística com as covariáveis selecionadas pelo *Stepwise*.

Com base na Figura 4.4, analisamos a métrica AUC (Área Sob a Curva) para avaliar o desempenho do modelo de classificação. Com um AUC de 0,81, concluímos que o modelo apresenta um bom desempenho na discriminação das classes da variável resposta. Além disso, observamos que ao aumentar a sensibilidade, ocorre uma redução na especificidade. Nesse sentido, buscando equilibrar sensibilidade e especificidade, empregamos as métricas Média-G e J, identificando um ponto de corte ótimo de 0,104 que otimiza ambas as métricas, o mesmo que foi obtido para a Regressão com todas covariáveis.

Para a *F-measure* baseada na Curva de Precisão-*Recall* construímos a seguinte Curva de Precisão-*Recall* na Figura 4.5.

Analisando a Figura 4.5, observamos que, embora alcancemos uma precisão mais elevada quando o *recall* é zero, ao atingirmos um *recall* superior a $0,125$, a precisão volta a aumentar. Dessa forma, a métrica F visa equilibrar precisão e *Recall*. Sendo assim, o valor que otimiza a função é $K = 0,001$.

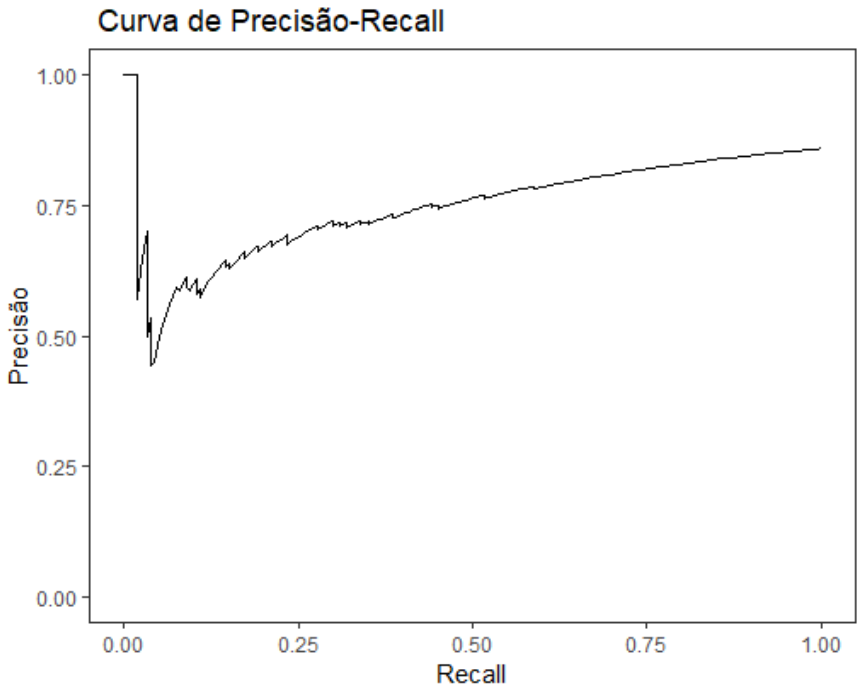


Figura 4.5: Curva de Precisão-*Recall* das predições do ajuste de Regressão Logística com as covariáveis selecionadas pelo *Stepwise*.

Desse modo, apresentamos na Figura 4.6 as três métricas utilizadas para obter os pontos de corte.

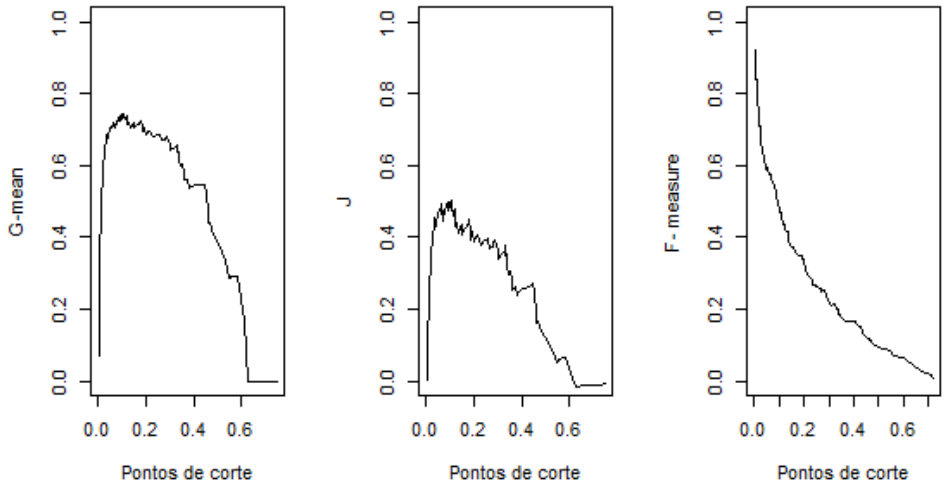


Figura 4.6: Pontos de corte para as métricas Média-G, J e *F-measure*, respectivamente.

Com respeito à Figura 4.6, temos o ponto que maximiza cada uma das métricas. Conforme mencionado anteriormente, a função de *F-measure* é maximizada em um local distinto das funções Média-G e J, evidenciado pelos diferentes formatos das curvas.

Após gerar os classificadores usando três pontos de corte distintos, elaboramos uma

matriz de confusão para comparar as proporções de acertos em todos os cenários, Tabela 4.4.

Tabela 4.4: Matriz de confusão para classificadores obtidos pela Regressão Logística com covariáveis selecionadas.

Valor Predito	**Valor Verdadeiro**					
	Ponto de corte = 0,5		Ponto de corte = 0,104		Ponto de corte = 0,001	
	Y = 0	Y = 1	Y = 0	Y = 1	Y = 0	Y = 1
Y = 0	0,83	0,12	0,56	0,02	0	0
Y = 1	0,03	0,02	0,3	0,12	0,86	0,14

Tabela 4.5: Medidas de Desempenho para os diferentes pontos de corte.

Medidas de Desempenho	**Ponto de Corte**		
	0,5	0,104	0,001
Acurácia	0,850	0,677	0,140
Sensibilidade (*Recall*)	0,143	0,857	1,000
Especificidade	0,967	0,648	0,000
Precisão	0,416	0,285	0,140
Valor Preditivo Negativo	0,873	0,965	0,000
Estatística F1	0,213	0,428	0,247

A partir da Tabela 4.4, observamos que, para o ponto de corte padrão de $0,5$, alcançamos uma taxa de acerto de $0,85$, no qual 83% dos indivíduos não diabéticos são corretamente classificados como não diabéticos, enquanto apenas 2% dos indivíduos diabéticos são classificados corretamente como diabéticos. Adicionalmente, 3% dos indivíduos não diabéticos são erroneamente classificados como diabéticos, e 12% dos indivíduos diabéticos são erroneamente classificados como não diabéticos. Para o ponto de corte de $0,104$, determinado pelas métricas Média-G e J, a taxa total de acertos é de $0,68$, destacando que 56% dos indivíduos não diabéticos são corretamente classificados, enquanto apenas 3% dos indivíduos não diabéticos são erroneamente classificados como diabéticos. Além disso, 12% dos indivíduos diabéticos são corretamente classificados, e 2% dos indivíduos diabéticos são erroneamente classificados como não diabéticos. Apesar da taxa total de acertos ser inferior em comparação com o ponto de corte de $0,5$, observamos uma sensibilidade mais elevada de $85,7\%$, conforme a Tabela 4.5. Por fim, ao considerar o ponto de corte de $0,001$, a taxa total de acertos é de $0,14$, sendo que apenas os indivíduos diabéticos são corretamente classificados (14%), enquanto os indivíduos não diabéticos são erroneamente classificados.

Com base na análise realizada e nas métricas de desempenho apresentadas na Tabela

4.5, notamos que, para o ponto de corte de $0,5$, as métricas de acurácia, especificidade e valor preditivo negativo são elevadas, o que era esperado, uma vez que a classificação da classe majoritária possui poucos erros. No entanto, a sensibilidade, precisão e Estatística F1 são significativamente baixas. Ao considerarmos o ponto de corte de $0,104$, observamos que suas métricas são mais proporcionais, apresentando uma sensibilidade razoavelmente alta, o que indica um melhor desempenho na identificação de indivíduos diabéticos, sem comprometer a taxa de acertos para indivíduos não diabéticos. Em última análise, para o ponto de corte de $0,001$, alcançamos o máximo de sensibilidade, pois todos os indivíduos diabéticos foram corretamente classificados. No entanto, essa escolha resulta em classificações incorretas para todos os indivíduos não diabéticos, levando a métricas como acurácia, especificidade, precisão e Estatística F1 muito baixas. Portanto, a conclusão é similar aquela descrita para os resultados da Regressão Logística utilizando todas as covariáveis.

4.3 Árvore de Classificação

Nesta seção, iremos aplicar o método de Árvore de Classificação previamente discutido na Seção 2.2. A aplicação da Árvore de Classificação foi implementada utilizando a linguagem de programação R (R Development Core Team, 2020).

A Árvore de Classificação obtida através da função *rpart* Therneau *et al.* (2015) é apresentada na Figura 4.7.

Ao interpretar a Figura 4.7, observamos numa visão geral o nó inicial Idade, os nós intermediários Dislipidemia, IMC, Idade e as folhas Atividade Física no Trabalho, TV (semana) e TV (final de semana). Além disso, podemos observar que o nó com maior representatividade de indivíduos com diabetes no conjunto de treinamento, 4% é o que segue a seguinte sequência: Idade maior ou igual do que 59 anos, possui dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, idade maior ou igual do que 65 anos e menor do que 69 anos.

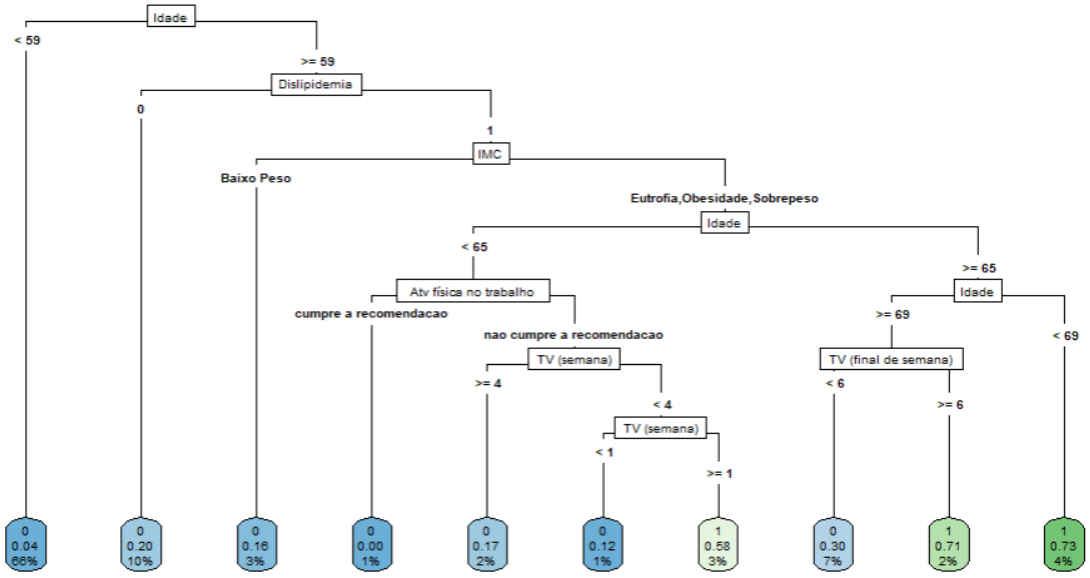


Figura 4.7: Árvore de Classificação.

Analisando cada uma das folhas temos que para a primeira folha, apenas é utilizada a variável idade Idade, sendo assim, 66% dos indivíduos do conjunto de treinamento possuem idade menor do que 59 anos com 4% de probabilidade de serem diabéticos, desse modo, indivíduos que estão nesse nó são classificados como não diabéticos. Por outro lado, o segundo nó possui a sequência de idade maior ou igual à 59 anos e não possuir dislipidemia, em que 10% dos indivíduos do conjunto de treinamento se enquadram nessa condição com 20% de probabilidade de serem diabéticos, desse modo, indivíduos que estão nesse nó são classificados como não diabéticos. Além disso, no terceiro nó temos a sequência idade maior ou igual à 59 anos, possuir dislipidemia e ser abaixo do peso, 3% dos indivíduos do conjunto de treinamento se encontram nessa categoria com 16% de probabilidade de serem diabéticos, então esses indivíduos são classificados como não diabéticos. Ademais, a quarta folha é composta pela seguinte sequência: idade maior ou igual à 59 anos, possuir dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, Idade menor de 65 e cumpre a recomendação de atividade física no trabalho, 1% dos indivíduos do conjunto de treinamento se encontram nessa categoria com 0% de probabilidade de serem diabéticos, logo esses indivíduos são classificados como não diabéticos. A quinta folha é composta pela seguinte sequência: idade maior ou igual à 59 anos, possuir dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, Idade menor de 65, não cumpre a recomendação de atividade física no trabalho e assiste TV por mais de 4 horas por um dia da semana, 2% dos indivíduos do conjunto de treinamento se encontram nessa categoria

com 17% de probabilidade de serem diabéticos, indivíduos que possuem essa sequência são classificados como não diabéticos.

Já a sexta folha, a sequência é dada por idade maior ou igual à 59 anos, possuir dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, Idade menor de 65, não cumpre a recomendação de atividade física no trabalho e assiste TV por menos de 1 hora por um dia da semana, sendo assim, 1% dos indivíduos do conjunto de treinamento possuem essa sequência com 12% de probabilidade de serem diabéticos, desse modo, indivíduos que estão nesse nó são classificados como não diabéticos. O sétimo nó possui a sequência de idade maior ou igual à 59 anos, possuir dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, Idade menor de 65, não cumpre a recomendação de atividade física no trabalho e assiste TV entre 1 hora e 4 horas por um dia da semana, em que 3% dos indivíduos do conjunto de treinamento se enquadram nessa condição com 58% de probabilidade de serem diabéticos, desse modo, indivíduos que estão nesse nó são classificados como diabéticos. Ademais, no oitavo nó temos a sequência idade maior ou igual à 59 anos, possuir dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, idade maior ou igual à 69 e assiste TV por menos de 6 horas em um dia do final de semana, 7% dos indivíduos do conjunto de treinamento se encontram nessa categoria com 30% de probabilidade de serem diabéticos, então esses indivíduos são classificados como não diabéticos. A nona folha é composta pela seguinte sequência: idade maior ou igual à 59 anos, possuir dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, idade maior ou igual à 69 e assiste TV entre por 6 horas ou mais em um dia do final de semana, 2% dos indivíduos do conjunto de treinamento se encontram nessa categoria com 71% de probabilidade de serem diabéticos, logo esses indivíduos são classificados como diabéticos. A última folha é composta pela seguinte sequência: idade maior ou igual à 59 anos, possuir dislipidemia, IMC igual à eutrofia, obesidade ou sobrepeso, idade maior ou igual do que 65 anos e menor do que 69 anos, 4% dos indivíduos do conjunto de treinamento se encontram nessa categoria com 73% de probabilidade de serem diabéticos, indivíduos que possuem essa sequência são classificados como diabéticos.

Após ajustar a Árvore de Classificação, realizamos as predições das observações do conjunto de validação e obtivemos as probabilidades $\hat{\mathbb{P}}(Y = 1|\boldsymbol{x})$. Dessa forma, como apresentado para Regressão Logística, ao desenvolver o classificador $g(\boldsymbol{x})$, optamos inicialmente por fixar o limiar K em $0,5$. Adicionalmente, devido ao desequilíbrio entre as classes, investigamos abordagens suplementares para a determinação de K, conforme

abordado na Seção 2.4.

Primeiramente, construímos a Curva ROC, apresentada na Figura 4.8.

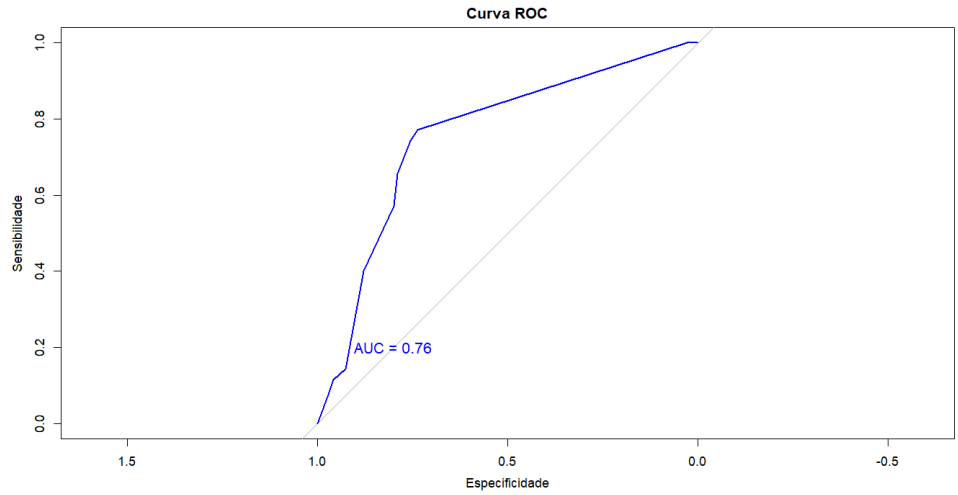


Figura 4.8: Curva ROC das predições do ajuste da Árvore de Classificação.

Analisando a Figura 4.8, possuímos um AUC de 0,76, o que indica que o modelo apresenta um bom desempenho na discriminação das classes da variável resposta,porém ainda assim é um pouco menor quando comparamos com as métricas AUC da Regressão Logística. Ademais, observamos que ao aumentar a sensibilidade, ocorre uma redução na especificidade. Nessa perspectiva, visando alcançar um equilíbrio entre sensibilidade e especificidade, utilizamos as métricas Média-G e J e identificamos um ponto de corte ideal de 0,083, o qual otimiza ambas as métricas.

Para a *F-measure* baseada na Curva de Precisão-*Recall* construímos a seguinte Curva de Precisão-*Recall* na Figura 4.9.

Analisando a Figura 4.9, percebemos que, mesmo obtendo uma precisão superior quando o *recall* é nulo, ao ultrapassarmos um *recall* superior a $0,25$, a precisão volta a se elevar. Assim, a métrica F é utilizada com o propósito de harmonizar a precisão e o *recall* e o valor que otimiza a função é $K = 0$.

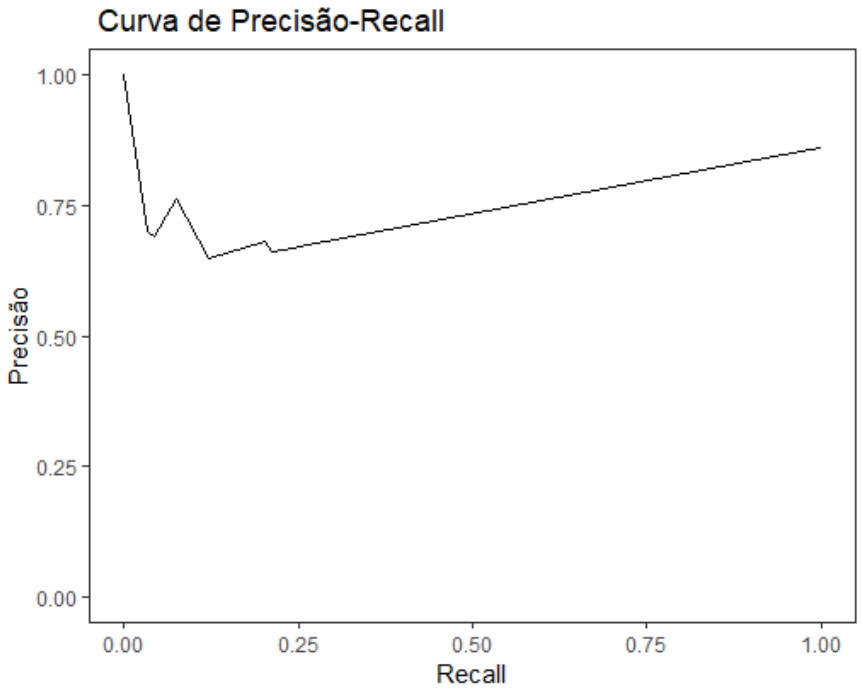


Figura 4.9: Curva de Precisão-*Recall* das predições do ajuste da Árvore de Classificação.

Sendo assim, apresentamos na Figura 4.10 as três métricas utilizadas para obter os pontos de corte.

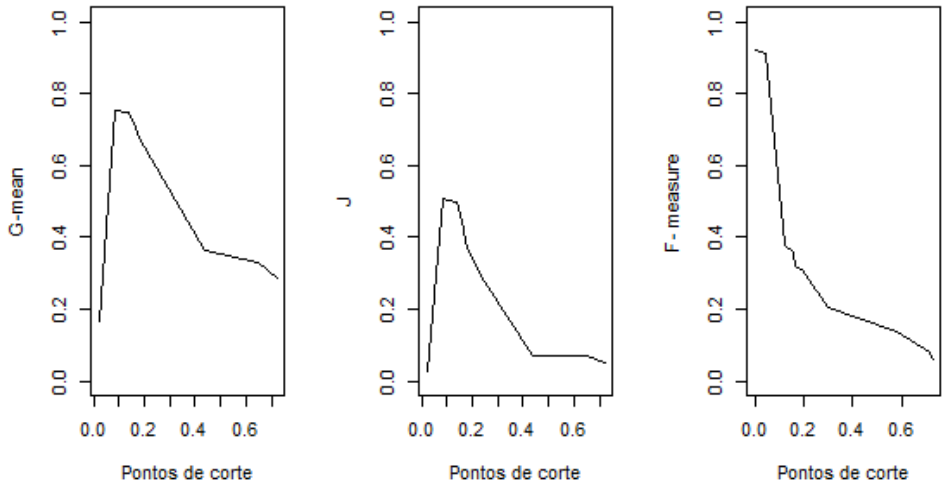


Figura 4.10: Pontos de corte para as métricas Média-G, J e *F-measure*, respectivamente.

Analisando a Figura 4.10, temos o ponto que maximiza cada uma das métricas. Conforme mencionado anteriormente, a função de *F-measure* é maximizada em um local distinto das funções Média-G e J, evidenciado pelos diferentes formatos das curvas.

Depois de criar os classificadores utilizando três pontos de corte diferentes, construímos

uma matriz de confusão na Tabela 4.6 para comparar as proporções de acertos em todos os cenários.

Tabela 4.6: Matriz de confusão para classificadores obtidos pela Árvore de Classificação com covariáveis selecionadas.

Valor Predito	**Valor Verdadeiro**					
	Ponto de corte = 0,5		Ponto de corte = 0,083		Ponto de corte = 0	
	Y = 0	Y = 1	Y = 0	Y = 1	Y = 0	Y = 1
Y = 0	0,79	0,12	0,63	0,03	0	0
Y = 1	0,06	0,02	0,23	0,11	0,86	0,14

Tabela 4.7: Medidas de Desempenho para os diferentes pontos de corte.

Medidas de Desempenho	**Ponto de Corte**		
	0,5	0,083	0
Acurácia	0,815	0,742	0,165
Sensibilidade (*Recall*)	0,143	0,771	1,000
Especificidade	0,925	0,737	0,028
Precisão	0,238	0,325	0,145
Valor Preditivo Negativo	0,867	0,951	1,000
Estatística F1	0,178	0,457	0,253

Analisando os dados apresentados na Tabela 4.6, para o ponto de corte padrão de $0,5$, alcançamos uma proporção total de acertos de $0,81$. Nesse cenário, 79% dos indivíduos não diabéticos são corretamente classificados como não diabéticos, enquanto apenas 2% dos indivíduos diabéticos são identificados corretamente. Adicionalmente, 6% dos indivíduos não diabéticos são erroneamente classificados como diabéticos, e 12% dos indivíduos diabéticos são erroneamente classificados como não diabéticos.

Ao adotar o ponto de corte de $0,083$, determinado pelas métricas G-Média e J, a proporção total de acertos é de $0,74$. Nota-se que 63% dos indivíduos não diabéticos são corretamente classificados, enquanto 23% dos indivíduos não diabéticos são erroneamente classificados como diabéticos. Além disso, 11% dos indivíduos diabéticos são corretamente classificados, e 3% dos indivíduos diabéticos são erroneamente classificados como não diabéticos. Apesar da proporção total de acertos ser inferior em comparação com o ponto de corte de $0,5$, observamos uma sensibilidade mais elevada de $77,1\%$ (conforme a Tabela 4.7).

Considerando o ponto de corte de 0, a proporção total de acertos é de $0,16$, em que a maior parte dos acertos são indivíduos diabéticos (14%), enquanto os indivíduos não diabéticos são erroneamente classificados em sua grande maioria (83%).

A análise das métricas de desempenho na Tabela 4.7 revela que, para o ponto de corte de $0,5$, as métricas de acurácia, especificidade e valor preditivo negativo são elevadas, refletindo a baixa incidência de erros na classificação da classe majoritária. Entretanto, a sensibilidade, precisão e Estatística F1 são notavelmente baixas. Ao adotarmos o ponto de corte de $0,083$, as métricas tornam-se mais equilibradas, destacando uma sensibilidade razoavelmente alta, indicando um melhor desempenho na identificação de indivíduos diabéticos sem comprometer a taxa de acertos para indivíduos não diabéticos. No entanto, para o ponto de corte de 0, a sensibilidade é maximizada, mas isso resulta em classificações incorretas para quase todos os indivíduos não diabéticos, levando a métricas como acurácia, especificidade, precisão e Estatística F1 significativamente baixas. Dessa forma, a conclusão é semelhante àquela descrita para os resultados da Regressão Logística utilizando todas as covariáveis e utilizando a seleção de variáveis *Stepwise*, em todos esses cenários, o ponto de corte obtido pelas métricas Média-G e J apresenta melhores resultados.

Capítulo 5

Considerações finais

Neste Trabalho de Conclusão de Curso com o objetivo de diagnosticar a diabetes em indivíduos utilizando algoritmos de classificação, primeiramente, revisamos os algoritmos de classificação, sendo eles Regressão Logística e Árvore de Classificação. Além disso, descrevemos medidas de desempenho para a avaliação da eficiência dos métodos para a classificação em estudo. Também foi apresentado a revisão da literatura sobre desbalanceamento de classes, pois as classes nem sempre são balanceadas e possuímos métricas como Média-G, J e *F-measure* que podem tratar esse desequilíbrio. Ademais, foi apresentada a estrutura do banco de dados composto por 829 indivíduos, uma variável resposta e 18 covariáveis.

Na análise descritiva e exploratória de dados foi estudado o comportamento da variável resposta individualmente e em relação a cada uma das covariáveis, onde pudemos observar, através dos gráficos, o comportamento das covariáveis em relação aos níveis da variável resposta, sendo eles, indivíduos diabéticos e não diabéticos.

Após isso, aplicamos os métodos de classificação para predizer se os indivíduos são diabéticos. Inicialmente, implementamos a Regressão Logística utilizando todas as covariáveis do nosso banco de dados, além disso, por causa do desbalanceamento de classes, construímos o classificador $g(\boldsymbol{x})$ consideramos os pontos de corte baseados nas métricas Média-G, J e *F-measure*, além do ponto de corte usual $0,5$. Analisando os resultados, notamos que o ponto de corte dado pelas métricas Média-G e J apresentava um melhor desempenho para o modelo, uma vez que além de termos uma sensibilidade alta, ou seja, maior identificação correta de indivíduos diabéticos dentre os indivíduos que realmente são diabéticos, houve também um equilíbrio melhor entre sensibilidade e especificidade, o que é essencial para o estudo, pois apesar de nosso interesse ser classificar corretamente

indivíduos diabéticos, também precisamos acertar o máximo possível de classificações de indivíduos não diabéticos.

Em seguida, implementamos a Regressão Logística utilizando apenas as covariáveis selecionadas pelo método *Stepwise* com técnica Bidirecional. As covariáveis selecionadas foram idade, IMC, hipertensão arterial, colesterol não-HDL e dislipidemia. Como na Regressão Logística utilizando todas as covariáveis, o melhor classificador $g(\boldsymbol{x})$ apresentado foi aquele que considerava os pontos de corte baseados nas métricas Média-G e J, pois além de uma sensibilidade alta, obtivemos um bom equilíbrio entre a sensibilidade e especificidade.

Logo após, construímos a Árvore de Classificação em que as variáveis selecionadas foram idade, dislipidemia, IMC, Prática de atividade física no trabalho, Horas assistindo TV durante um dia da semana e Horas assistindo TV durante um dia do final de semana. Assim como nos métodos apresentados anteriormente, o classificador $g(\boldsymbol{x})$ possui melhor desempenho quando considera os pontos de corte baseados nas métricas Média-G e J.

Por fim, para atender nosso objetivo principal de diagnosticar a diabetes em indivíduos, os achados pela Regressão Logística são consistentes aos da classificação por Árvore, porém o método de Árvore de Classificação utilizando o ponto de corte dado pelas métricas Média-G e J, apresenta uma sutil melhora nas métricas Acurácia, Especificidade, Precisão e Estatística $F1$, além de possuir um melhor equilíbrio entre a sensibilidade e especificidade, o que é muito importante na área da saúde e para este estudo em questão.

Referências Bibliográficas

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

Hassan, M. M. e Amiri, N. (2019). Classification of imbalanced data of diabetes disease using machine learning algorithms. *Age (years)*, **21**(81), 33–24.

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.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sisodia, D. e Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, **132**, 1578–1585.

Therneau, T., Atkinson, B., Ripley, B. e Ripley, M. B. (2015). Package 'rpart'. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*.

Apêndice A

Códigos Computacionais

Para a análise, o seguinte código, implementado no programa computacional R (R Development Core Team, 2020), foi utilizado.

```
#Pacotes
library(ggplot2)
library(colorspace)
library(dplyr)

#base de dados
dados_diabetes <- data.table::fread('C:/Users/larissa/OneDrive/Documentos/Larissa/ufscar/TG/Banco_Andressa_DM.csv',
                           sep = ";",encoding = "UTF-8")

View(dados_diabetes)

#Transformar celulas vazias em NA
unique(dados_diabetes$imc_cat)
dados_diabetes$imc_cat [dados_diabetes$imc_cat == ""] <- NA
unique(dados_diabetes$imc_cat)

unique(dados_diabetes$fumo2)
dados_diabetes$fumo2 [dados_diabetes$fumo2 == ""] <- NA

unique(dados_diabetes$alcool)
dados_diabetes$alcool [dados_diabetes$alcool == ""] <- NA

unique(dados_diabetes$AF_LAZ_cat_OMS_1)
dados_diabetes$AF_LAZ_cat_OMS_1 [dados_diabetes$AF_LAZ_cat_OMS_1 == ""] <- NA

unique(dados_diabetes$AF_DOM_cat_OMS_1)
dados_diabetes$AF_DOM_cat_OMS_1 [dados_diabetes$AF_DOM_cat_OMS_1 == ""] <- NA

unique(dados_diabetes$AF_TRANSP_cat_OMS_1)
dados_diabetes$AF_TRANSP_cat_OMS_1 [dados_diabetes$AF_TRANSP_cat_OMS_1 == ""] <- NA
```

```
unique(dados_diabetes$AF_TRAB_cat_OMS_1)
dados_diabetes$AF_TRAB_cat_OMS_1 [dados_diabetes$AF_TRAB_cat_OMS_1 == ""] <- NA


#ver NAs em cada uma das variaveis
sapply(dados_diabetes, function(x) sum(is.na(x)))

#Graficos das variaveis

dados_diabetes$diabetes2 <- sub("0", "Não", dados_diabetes$diabetes2)
dados_diabetes$diabetes2 <- sub("1", "Sim", dados_diabetes$diabetes2)

y<-table(dados_diabetes$diabetes2)
y <- prop.table(y)
y<- as.data.frame(y)

colnames(y) <- c("Diabetes", "Frequência Relativa")

ggplot(y, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(stat="identity", fill= "deepskyblue4") + theme_classic(base_size = 11)+ ylim(0,0.9)


#Base de dados desconsiderando adolescentes
adolescente <- which( dados_diabetes$faixa_etaria==0)

dados_s_adoles<- dados_diabetes[-adolescente,]

#Gráfico Variável respostas desconsiderando adolescentes
y<-table(dados_s_adoles$diabetes2)
y <- prop.table(y)
y<- as.data.frame(y)

colnames(y) <- c("Diabetes", "Frequência Relativa")

ggplot(y, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(stat="identity", fill= "deepskyblue4") +  theme_classic(base_size = 11)+ ylim(0,0.9)


#Gráficos covariáveis

#Sexo
valores_unicos_sexo <- unique(dados_diabetes$sexo)
dados_diabetes$sexo <- sub("feminino", "Feminino", dados_diabetes$sexo)
dados_diabetes$sexo <- sub("masculino", "Masculino", dados_diabetes$sexo)


sexo<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$sexo)
sexo <- prop.table(sexo,1)
sexo2<- as.data.frame(sexo)
colnames(sexo2) <- c("Diabetes", "Sexo", "Frequência Relativa")
```

```
ggplot(sexo2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =Sexo), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +  theme_classic(base_size = 11)+ ylim(0,0.70)


#Faixa etária

valores_unicos_faixa_etaria <- unique(dados_diabetes$faixa_etaria)
dados_diabetes$faixa_etaria <- sub("2", "Idosos", dados_diabetes$faixa_etaria)
dados_diabetes$faixa_etaria <- sub("1", "Adultos", dados_diabetes$faixa_etaria)
dados_diabetes$faixa_etaria <- sub("0", "Adolescentes", dados_diabetes$faixa_etaria)


fx_etaria<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$faixa_etaria)
fx_etaria <- prop.table(fx_etaria,1)
fx_etaria2<- as.data.frame(fx_etaria)
colnames(fx_etaria2) <- c("Diabetes", "Faixa Etária", "Frequência Relativa")

ggplot(fx_etaria2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Faixa Etária`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +  theme_classic(base_size = 11)


#IMC

imc<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$imc_cat)
imc <- prop.table(imc,1)
imc2<- as.data.frame(imc)
colnames(imc2) <- c("Diabetes", "IMC", "Frequência Relativa")


ggplot(imc2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =IMC), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +  theme_classic(base_size = 11)


#Hipertensão Arterial
valores_unicos_has <- unique(dados_diabetes$has2)
dados_diabetes$has2 <- sub("0", "Não", dados_diabetes$has2)
dados_diabetes$has2 <- sub("1", "Sim", dados_diabetes$has2)


has<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$has2)
has <- prop.table(has,1)
has2<- as.data.frame(has)
colnames(has2) <- c("Diabetes", "Hipertensão Arterial", "Frequência Relativa")
```

```
ggplot(has2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Hipertensão Arterial`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +  theme_classic(base_size = 11) + ylim(0,0.85)

###desconsiderando adolescentes

valores_unicos_has <- unique(dados_s_adoles$has2)
dados_s_adoles$has2 <- sub("0", "Não", dados_s_adoles$has2)
dados_s_adoles$has2 <- sub("1", "Sim", dados_s_adoles$has2)


has<- table(as.character(dados_s_adoles$diabetes2),dados_s_adoles$has2)
has <- prop.table(has,1)
has2<- as.data.frame(has)
colnames(has2) <- c("Diabetes", "Hipertensão Arterial", "Frequência Relativa")

ggplot(has2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Hipertensão Arterial`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +  theme_classic(base_size = 11) + ylim(0,0.85)



#Colesterol não HDL

valores_unicos_col <- unique(dados_diabetes$colnaohdl_cat)

dados_diabetes$colnaohdl_cat <- sub("0", "Adequado", dados_diabetes$colnaohdl_cat)
dados_diabetes$colnaohdl_cat <- sub("1", "Inadequado", dados_diabetes$colnaohdl_cat)


coles<- table(as.character(dados_diabetes$diabetes2),as.character(dados_diabetes$colnaohdl_cat))
coles <- prop.table(coles,1)
coles2<- as.data.frame(coles)
colnames(coles2) <- c("Diabetes", "Colesterol não-HDL", "Frequência Relativa")

ggplot(coles2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill = `Colesterol não-HDL`),position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +  theme_classic(base_size = 11)


#colesterol HDL


valores_unicos_hdl <- unique(dados_diabetes$hdl_baixo)
dados_diabetes$hdl_baixo <- sub("0", "Adequado", dados_diabetes$hdl_baixo)
dados_diabetes$hdl_baixo <- sub("1", "Inadequado", dados_diabetes$hdl_baixo)
```

```
hdl<- table(as.character(dados_diabetes$diabetes2),as.character(dados_diabetes$hdl_baixo))
hdl <- prop.table(hdl,1)
hdl2<- as.data.frame(hdl)
colnames(hdl2) <- c("Diabetes", "Colesterol HDL", "Frequência Relativa")

ggplot(hdl2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill = `Colesterol HDL`),position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +  theme_classic(base_size = 11) +ylim(0,0.80)


#Dislipidemia
valores_unicos_dlp2 <- unique(dados_diabetes$dlp2)
dados_diabetes$dlp2 <- sub("0", "Não", dados_diabetes$dlp2)
dados_diabetes$dlp2 <- sub("1", "Sim", dados_diabetes$dlp2)


dislipi<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$dlp2)
dislipi <- prop.table(dislipi,1)
dislipi2<- as.data.frame(dislipi)
colnames(dislipi2) <- c("Diabetes", "Dislipidemia", "Frequência Relativa")

ggplot(dislipi2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =Dislipidemia), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11)


#doença mental
valores_unicos_C221A <- unique(dados_diabetes$C221A)
dados_diabetes$C221A <- sub("1", "Não", dados_diabetes$C221A)
dados_diabetes$C221A <- sub("2", "Sim", dados_diabetes$C221A)

diabetes<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$C221A)
diabetes <- prop.table(diabetes,1)
diabetes2<- as.data.frame(diabetes)
colnames(diabetes2) <- c("Diabetes", "Problema Emocional", "Frequência Relativa")

ggplot(diabetes2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Problema Emocional`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11)



#Fumo

valores_unicos_fumo2 <- unique(dados_diabetes$fumo2)
```

```
dados_diabetes$fumo2 <- sub("nunca", "Nunca fumou", dados_diabetes$fumo2)
dados_diabetes$fumo2 <- sub("ex-fumante e fumante", "Ex fumante ou Fumante", dados_diabetes$fumo2)

fumo<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$fumo2)
fumo <- prop.table(fumo,1)
fumo2<- as.data.frame(fumo)
colnames(fumo2) <- c("Diabetes", "Tabagismo", "Frequência Relativa")


ggplot(fumo2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =Tabagismo), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11) +ylim(0,0.8)

###desconsiderando adolescentes
valores_unicos_fumo2 <- unique(dados_s_adoles$fumo2)
dados_s_adoles$fumo2 <- sub("nunca", "Nunca fumou", dados_s_adoles$fumo2)
dados_s_adoles$fumo2 <- sub("ex-fumante e fumante", "Ex fumante ou Fumante", dados_s_adoles$fumo2)

fumo<- table(as.character(dados_s_adoles$diabetes2),dados_s_adoles$fumo2)
fumo <- prop.table(fumo,1)
fumo2<- as.data.frame(fumo)
colnames(fumo2) <- c("Diabetes", "Tabagismo", "Frequência Relativa")


ggplot(fumo2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =Tabagismo), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11) +ylim(0,0.8)



#Alcool


valores_unicos <- unique(dados_diabetes$alcool)
dados_diabetes$alcool <- sub("parou de beber", "Parou de beber", dados_diabetes$alcool)
dados_diabetes$alcool <- sub("nunca", "Nunca bebeu", dados_diabetes$alcool)
dados_diabetes$alcool <- sub("bebe atualmente", "Bebe atualmente", dados_diabetes$alcool)

alcool<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$alcool)
alcool <- prop.table(alcool,1)
alcool2<- as.data.frame(alcool)
colnames(alcool2) <- c("Diabetes", "Consumo de bebida alcoólica", "Frequência Relativa")


ggplot(alcool2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Consumo de bebida alcoólica`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11) +ylim(0,0.7)
```

```
### desconsiderando adolescentes
valores_unicos <- unique(dados_s_adoles$alcool)
dados_s_adoles$alcool <- sub("parou de beber", "Parou de beber", dados_s_adoles$alcool)
dados_s_adoles$alcool <- sub("nunca", "Nunca bebeu", dados_s_adoles$alcool)
dados_s_adoles$alcool <- sub("bebe atualmente", "Bebe atualmente", dados_s_adoles$alcool)

alcool<- table(as.character(dados_s_adoles$diabetes2),dados_s_adoles$alcool)
alcool <- prop.table(alcool,1)
alcool2<- as.data.frame(alcool)
colnames(alcool2) <- c("Diabetes", "Consumo de bebida alcoólica", "Frequência Relativa")


ggplot(alcool2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Consumo de bebida alcoólica`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  +theme_classic(base_size = 11) +ylim(0,0.7)



#atividada física no lazer

valores_unicos <- unique(dados_diabetes$AF_LAZ_cat_OMS_1)
dados_diabetes$AF_LAZ_cat_OMS_1 <-
sub("nao cumpre a recomendacao", "Não cumpre a recomendação", dados_diabetes$AF_LAZ_cat_OMS_1)
dados_diabetes$AF_LAZ_cat_OMS_1 <-
sub("cumpre a recomendacao", "Cumpre a recomendação", dados_diabetes$AF_LAZ_cat_OMS_1)


atividade<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$AF_LAZ_cat_OMS_1)
atividade <- prop.table(atividade,1)
atividade2<- as.data.frame(atividade)
colnames(atividade2) <- c("Diabetes", "Atividade física no lazer", "Frequência Relativa")



ggplot(atividade2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Atividade física no lazer`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11) +ylim(0,0.87)


#atividada física domestica

valores_unicos <- unique(dados_diabetes$AF_DOM_cat_OMS_1)
dados_diabetes$AF_DOM_cat_OMS_1 <-
sub("nao cumpre a recomendacao", "Não cumpre a recomendação", dados_diabetes$AF_DOM_cat_OMS_1)
dados_diabetes$AF_DOM_cat_OMS_1 <-
sub("cumpre a recomendacao", "Cumpre a recomendação", dados_diabetes$AF_DOM_cat_OMS_1)
```

```
atividade<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$AF_DOM_cat_OMS_1)
atividade <- prop.table(atividade,1)
atividade2<- as.data.frame(atividade)
colnames(atividade2) <- c("Diabetes", "Atividade física doméstica", "Frequência Relativa")


ggplot(atividade2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Atividade física doméstica`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11) + ylim(0,0.87)




#atividada física no transporte

valores_unicos <- unique(dados_diabetes$AF_TRANSP_cat_OMS_1)
dados_diabetes$AF_TRANSP_cat_OMS_1 <-
sub("nao cumpre a recomendacao", "Não cumpre a recomendação", dados_diabetes$AF_TRANSP_cat_OMS_1)
dados_diabetes$AF_TRANSP_cat_OMS_1 <-
sub("cumpre a recomendacao", "Cumpre a recomendação", dados_diabetes$AF_TRANSP_cat_OMS_1)


atividade<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$AF_TRANSP_cat_OMS_1)
atividade <- prop.table(atividade,1)
atividade2<- as.data.frame(atividade)
colnames(atividade2) <- c("Diabetes", "Atividade física no transporte", "Frequência Relativa")



ggplot(atividade2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Atividade física no transporte`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11)


#atividada física no trabalho

valores_unicos <- unique(dados_diabetes$AF_TRAB_cat_OMS_1)
dados_diabetes$AF_TRAB_cat_OMS_1 <-
sub("nao cumpre a recomendacao", "Não cumpre a recomendação", dados_diabetes$AF_TRAB_cat_OMS_1)
dados_diabetes$AF_TRAB_cat_OMS_1 <-
sub("cumpre a recomendacao", "Cumpre a recomendação", dados_diabetes$AF_TRAB_cat_OMS_1)


atividade<- table(as.character(dados_diabetes$diabetes2),dados_diabetes$AF_TRAB_cat_OMS_1)
atividade <- prop.table(atividade,1)
atividade2<- as.data.frame(atividade)
colnames(atividade2) <- c("Diabetes", "Atividade física no trabalho", "Frequência Relativa")
```

```
ggplot(atividade2, aes(y=`Frequência Relativa`, x=Diabetes)) +
  geom_bar(aes(fill =`Atividade física no trabalho`), position = "dodge", stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+ scale_fill_discrete_sequential(palette = "Teal")
  + theme_classic(base_size = 11)


#Variáveis contínuas


#idade
ggplot(data=dados_diabetes, aes(x=diabetes2, y=idade, fill= diabetes2))+ geom_boxplot(show.legend = F, alpha = .9) +
  theme_classic(base_size = 11) + scale_fill_discrete_sequential(palette = "Teal") +
  xlab("Diabetes") +
  ylab("Idade") +ylim(10,95)

summary(dados_diabetes$idade)

### desconsiderando adolescentes
ggplot(data=dados_s_adoles, aes(x=diabetes2, y=idade, fill= diabetes2))+ geom_boxplot(show.legend = F, alpha = .9) +
  theme_classic(base_size = 11) + scale_fill_discrete_sequential(palette = "Teal") +
  xlab("Diabetes") +
  ylab("Idade") +ylim(10,95)

summary(dados_s_adoles$idade)

#Horas assistindo tv durante um dia da semana

unique(dados_diabetes$k206_hrs)

tv_s<- data.frame(dados_diabetes$diabetes2,dados_diabetes$k206_hrs)
colnames(tv_s) <- c("Diabetes", "k206_hrs")


which(is.na(tv_s$k206_hrs)==T)

tv_s <- tv_s[-c(50,97,242,642,869,685),]
unique(tv_s$k206_hrs)

ggplot(data = tv_s, aes(x = k206_hrs, fill = Diabetes, y = ..density..)) +
  geom_histogram(position = "identity", show.legend = TRUE, alpha = 0.5,bins=20) +
  theme_classic(base_size = 11) +
  scale_fill_discrete_sequential(palette = "Teal") +
  xlab("Tempo assistindo TV durante um dia da semana") +
  ylab("Densidade")



#Horas assistindo tv durante um dia do final de semana
```

```
unique(dados_diabetes$k207_hrs)

tv_fds<- data.frame(dados_diabetes$diabetes2,dados_diabetes$k207_hrs)
colnames(tv_fds) <- c("Diabetes", "k207_hrs")


which(is.na(tv_fds$k207_hrs)==T)

tv_fds <- tv_fds[-c(50,97,242,642,869,197,325,664,749),]
unique(tv_fds$k207_hrs)

ggplot(data = tv_fds, aes(x = k207_hrs, fill = Diabetes, y = ..density..)) +
  geom_histogram(position = "identity", show.legend = TRUE, alpha = 0.5,bins=20) +
  theme_classic(base_size = 11) +
  scale_fill_discrete_sequential(palette = "Teal") +
  xlab("Tempo assistindo TV durante um dia do final de semana") +
  ylab("Densidade")



#Horas no computador durante um dia da semana

unique(dados_diabetes$k208_hrs)

pc_s<- data.frame(dados_diabetes$diabetes2,dados_diabetes$k208_hrs)
colnames(pc_s) <- c("Diabetes", "k208_hrs")


which(is.na(pc_s$k208_hrs)==T)

pc_s <- pc_s[-c(50,97,242,474,642,869,53),]
unique(pc_s$k208_hrs)

ggplot(data = pc_s, aes(x = k208_hrs, fill = Diabetes, y = ..density..)) +
  geom_histogram(position = "identity", show.legend = T, alpha = 0.5,bins=20) +
  theme_classic(base_size = 11) +
  scale_fill_discrete_sequential(palette = "Teal") +
  xlab("Tempo no computador durante um dia da semana") +
  ylab("Densidade")


#Horas no computador durante um dia do final de semana

unique(dados_diabetes$k209_hrs)

pc_fds<- data.frame(dados_diabetes$diabetes2,dados_diabetes$k209_hrs)
colnames(pc_fds) <- c("Diabetes", "k209_hrs")
```

```
which(is.na(pc_fds$k209_hrs)==T)

pc_fds <- pc_fds[-c(50,97,242,474,642,869),]
unique(pc_fds$k209_hrs)

ggplot(data = pc_fds, aes(x = k209_hrs, fill = Diabetes, y = ..density..)) +
  geom_histogram(position = "identity", show.legend = T, alpha = 0.5,bins=20) +
  theme_classic(base_size = 11) +
  scale_fill_discrete_sequential(palette = "Teal") +
  xlab("Tempo no computador durante um dia do final de semana") +
  ylab("Densidade")

#Correlação

dados_diabetes <- na.omit(dados_diabetes)
dim(dados_diabetes)

#Tranformando variáveis em fator
dados_diabetes$faixa_etaria <- as.factor(dados_diabetes$faixa_etaria)
dados_diabetes$diabetes2 <- as.factor(dados_diabetes$diabetes2)
dados_diabetes$has2 <- as.factor(dados_diabetes$has2)
dados_diabetes$colnaohdl_cat <- as.factor(dados_diabetes$colnaohdl_cat)
dados_diabetes$hdl_baixo <- as.factor(dados_diabetes$hdl_baixo)
dados_diabetes$dlp2 <- as.factor(dados_diabetes$dlp2)
dados_diabetes$C221A <- as.factor(dados_diabetes$C221A)
glimpse(dados_diabetes)

#Correlação entre variáveis contínuas

x<- data.frame(dados_diabetes$k206_hrs,dados_diabetes$k207_hrs, dados_diabetes$k208_hrs, dados_diabetes$k209_hrs,
dados_diabetes$idade)
names(x)[names(x) == 'dados_diabetes.k206_hrs'] <- 'TV (semana)'
names(x)[names(x) == 'dados_diabetes.k207_hrs'] <- 'TV (final de semana)'
names(x)[names(x) == 'dados_diabetes.k208_hrs'] <- 'Computador (semana)'
names(x)[names(x) == 'dados_diabetes.k209_hrs'] <- 'Computador (final de semana)'
names(x)[names(x) == 'dados_diabetes.idade'] <- 'Idade'

matriz_corr <- cor(x,use = "complete.obs")
corrplot(matriz_corr,method = "color", tl.col = "#424242", tl.srt = 45,
         addCoef.col = "#ffffff", tl.cex = 0.8, number.cex = 0.9, number.font = 1,
         cl.cex = 0.9,
         col = colorRampPalette(c("lightblue1","turquoise4","darkslategray"))(200))


#Divisão entre treino e teste

set.seed(1875)
indices <- createDataPartition(y = dados_diabetes$diabetes2, p = 0.7, list = FALSE)
base_treino <- dados_diabetes[indices, ]  # 70% dos dados para treinamento
base_teste <- dados_diabetes[-indices, ]  # 30% dos dados para teste
```

```
y_treino<-table(base_treino$diabetes2)
y_treino <- prop.table(y_treino)

y_teste<-table(base_teste$diabetes2)
y_teste <- prop.table(y_teste)


#Regressão Logística

modelo <- glm(diabetes2 ~ sexo +idade + imc_cat + has2 + colnaohdl_cat +
                hdl_baixo + dlp2 + C221A +k206_hrs + k207_hrs + k208_hrs +k209_hrs
                + fumo2 + alcool + AF_LAZ_cat_OMS_1 + AF_DOM_cat_OMS_1+
                AF_TRANSP_cat_OMS_1 + AF_TRAB_cat_OMS_1, data= base_treino, family = binomial(link = "logit"))
summary(modelo)

#coeficientes do modelo
exp(coef(modelo))

#previsão no teste
previsoes <- predict(modelo, newdata = base_teste[,-5], type= "response")

y_teste <- unlist(base_teste[,5])

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

# Plotar a curva ROC
plot(roc_curve, main = "Curva ROC", col = "blue", lwd = 2, ylab="Sensibilidade", xlab= "Especificidade")
auc_value <- auc(roc_curve)
text(0.8, 0.2, paste("AUC =", round(auc_value, 2)), col = "blue", cex = 1.2)

#g-mean

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

#estatístia J

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

#Matriz de Confusão

previsoes_binarias <- ifelse(previsoes > g, '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")

#Curva precisão-Recall
data_teste <- data.frame(y_teste,previsoes)
curve <- pr_curve(data_teste, truth= y_teste, previsoes)

autoplot(curve)+ theme(panel.grid = element_blank()) + labs(title = " Curva de Precisão-Recall",
                                                            x = "Recall",
                                                            y = "Precisão")

Recall<-curve$recall
Precision<- curve$precision
F_Measure = (2 * Precision * Recall) / (Precision + Recall)
max_F_Measure<- max(F_Measure)
id <- which(F_Measure==max_F_Measure)
curve$.threshold[id]

previsoes_binarias3 <- ifelse(previsoes > curve$.threshold[id], '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias3),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")

#Comparação curvas
par(mfrow=c(1,3))
plot(roc_curve$thresholds,g_mean,xlab="Pontos de corte",
     ylab="G-mean",type="l",ylim=c(0,1))
plot(roc_curve$thresholds,j_estat,xlab="Pontos de corte",
     ylab="J",type="l",ylim=c(0,1))
plot(curve$.threshold,F_Measure,xlab="Pontos de corte",
     ylab="F- measure",type="l",ylim=c(0,1))


#Regressão Logística utilizando Stepwise

step(modelo)

modelo1<- glm(diabetes2 ~ idade + imc_cat + has2 + colnaohdl_cat + dlp2, data= base_treino,
family = binomial(link = "logit"))
```

```
summary(modelo1)

#coeficientes do modelo
exp(coef(modelo1))

#previsão no teste
previsoes1 <- predict(modelo1, newdata = base_teste[,-5], type= "response")



#Curva ROC
roc_curve1 <- roc(response= y_teste, predictor= previsoes1)

# Plotar a curva ROC
plot(roc_curve1, main = "Curva ROC", col = "blue", lwd = 2, ylab="Sensibilidade", xlab= "Especificidade")
auc_value <- auc(roc_curve)
text(0.8, 0.2, paste("AUC =", round(auc_value, 2)), col = "blue", cex = 1.2)

#g-mean

g_mean <- sqrt(roc_curve1$sensitivities*roc_curve1$specificities)
max_g_mean<- max(g_mean)
id <- which(g_mean==max_g_mean)
g <- roc_curve1$thresholds[id]

#estatístia J

j_estat <- roc_curve1$sensitivities+roc_curve1$specificities-1
max_j_estat<- max(j_estat)
id <- which(j_estat==max_j_estat)
j <- roc_curve1$thresholds[id]

#Matriz de Confusão

previsoes_binarias <- ifelse(previsoes1 > g, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")



previsoes_binarias2 <- ifelse(previsoes1 > 0.5, '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias2),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")

#Curva precisão-Recall
data_teste <- data.frame(y_teste,previsoes1)
curve <- pr_curve(data_teste, truth= y_teste, previsoes1)
```

```
autoplot(curve)+ theme(panel.grid = element_blank()) + labs(title = " Curva de Precisão-Recall",
                                                            x = "Recall",
                                                            y = "Precisão")

Recall<-curve$recall
Precision<- curve$precision
F_Measure = (2 * Precision * Recall) / (Precision + Recall)
max_F_Measure<- max(F_Measure)
id <- which(F_Measure==max_F_Measure)
curve$.threshold[id]

previsoes_binarias3 <- ifelse(previsoes1 > curve$.threshold[id], '1', '0')
confusionMatrix(data = as.factor(previsoes_binarias3),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")

#Comparação curvas
par(mfrow=c(1,3))
plot(roc_curve1$thresholds,g_mean,xlab="Pontos de corte",
     ylab="G-mean",type="l",ylim=c(0,1))
plot(roc_curve1$thresholds,j_estat,xlab="Pontos de corte",
     ylab="J",type="l",ylim=c(0,1))
plot(curve$.threshold,F_Measure,xlab="Pontos de corte",
     ylab="F- measure",type="l",ylim=c(0,1))

#Árvores de Classificação

library(rpart)
library(rpart.plot)

treinamento_2 <- data.frame(base_treino)
y_treino<- treinamento_2$diabetes2

#names(treinamento_2)[names(treinamento_2) == 'idade'] <- 'Idade'
#names(treinamento_2)[names(treinamento_2) == 'dlp2'] <- 'Dislipidemia'
#names(treinamento_2)[names(treinamento_2) == 'imc_cat'] <- 'IMC'
#names(treinamento_2)[names(treinamento_2) == 'AF_TRAB_cat_OMS_1'] <- 'Atv física no trabalho'
#names(treinamento_2)[names(treinamento_2) == 'k206_hrs'] <- 'TV (semana)'
#names(treinamento_2)[names(treinamento_2) == 'k207_hrs'] <- 'TV (final de semana)'



par(mfrow=c(1,1))
# ajuste da arvore
arvore <- rpart(y_treino ~.,data = treinamento_2[,-5],method = "class")

rpart.plot(arvore, type = 5,cex=0.5)



# poda
```

```
#melhorCp <- arvore$cptable[which.min(arvore$cptable[,"xerror"]),
                            "CP"]
#poda <- prune(arvore, cp = melhorCp)

# arvore
#rpart.plot(poda, type = 5)

#predito arvore
predito_arvore <- predict(arvore,data.frame(base_teste[,-5]))

previsoes_binarias1 <- ifelse(predito_arvore[,2] > 0.5, '1', '0')

confusionMatrix(data = as.factor(previsoes_binarias1),
                reference = as.factor(y_teste),
                positive = '1',
                mode = "everything")

#Curva ROC

y_teste <- base_teste$diabetes2
roc_curve <- roc(response=y_teste, predictor= predito_arvore[,2])

# Plotar a curva ROC
plot(roc_curve, main = "Curva ROC", col = "blue", lwd = 2, ylab="Sensibilidade", xlab= "Especificidade")
auc_value <- auc(roc_curve)
text(0.8, 0.2, paste("AUC =", round(auc_value, 2)), col = "blue", 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 <- roc_curve$thresholds[id]

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

#Matriz de Confusão

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

predito <- as.double(predito_arvore[,2])
data_teste <- data.frame(y_teste,predito)
```

```
curve <- pr_curve(data_teste, truth= y_teste, predito)

autoplot(curve)+ theme(panel.grid = element_blank()) + labs(title = " Curva de Precisão-Recall",
       x = "Recall",
       y = "Precisão")


Recall<-curve$recall
Precision<- curve$precision
F_Measure = (2 * Precision * Recall) / (Precision + Recall)
max_F_Measure<- max(F_Measure)
id <- which(F_Measure==max_F_Measure)
curve$.threshold[id]


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

#Comparação curvas
par(mfrow=c(1,3))
plot(roc_curve$thresholds,g_mean,xlab="Pontos de corte",
     ylab="G-mean",type="l",ylim=c(0,1))
plot(roc_curve$thresholds,j_estat,xlab="Pontos de corte",
     ylab="J",type="l",ylim=c(0,1))
plot(curve$.threshold,F_Measure,xlab="Pontos de corte",
     ylab="F- measure",type="l",ylim=c(0,1))
```