UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus Presidente Prudente

WILLIAM YUJI KOHATSU TANIGAVA

ESTUDO COMPARATIVO DE MODELOS DE CLASSIFICAÇÃO QUANTO AO DESFECHO DE INTERNAÇÕES HOSPITALARES POR INFARTO AGUDO DO MIOCÁRDIO NO MUNICÍPIO DE SÃO PAULO ENTRE 2012 A 2022

PRESIDENTE PRUDENTE
2024

WILLIAM YUJI KOHATSU TANIGAVA

ESTUDO COMPARATIVO DE MODELOS DE CLASSIFICAÇÃO QUANTO AO DESFECHO DE INTERNAÇÕES HOSPITALARES POR INFARTO AGUDO DO MIOCÁRDIO NO MUNICÍPIO DE SÃO PAULO ENTRE 2012 A 2022

Relatório Final do Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp.
Orientador: Prof. Dr. Mário Hissamitsu Tarumoto

PRESIDENTE PRUDENTE
2024

T164e

Tanigava, William Yuji Kohatsu

Estudo comparativo de modelos de classificação quanto ao desfecho de internações hospitalares por infarto agudo do miocárdio no município de São Paulo entre 2012 a 2022 / William Yuji Kohatsu Tanigava. -- Presidente Prudente, 2024

57 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientador: Mário Hissamitsu Tarumoto

1. Árvores de decisão. 2. Regressão logística. 3. Floresta aleatória. 4. Infarto do miocárdio. 5. DATASUS. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

WILLIAM YUJI KOHATSU TANIGAVA

ESTUDO COMPARATIVO DE MODELOS DE CLASSIFICAÇÃO QUANTO AO DESFECHO DE INTERNAÇÕES HOSPITALARES POR INFARTO AGUDO DO MIOCÁRDIO NO MUNICÍPIO DE SÃO PAULO ENTRE 2012 A 2022

Relatório Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do Curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Documento assinado digitalmente
gov.br
MARIO HISSAMITSU TARUMOTO
Data: 17/12/2024 18:10:39-0300
Verifique em https://validar.iti.gov.br

Orientador: _____

Prof. Dr. Mário Hissamitsu Tarumoto

Departamento de Estatística

Documento assinado digitalmente
gov.br
MIRIAM RODRIGUES SILVESTRE
Data: 17/12/2024 14:17:00-0300
Verifique em https://validar.iti.gov.br

_____

Profª. Dra. Miriam Rodrigues Silvestre

Departamento de Estatística

Presidente Prudente, 07 de dezembro de 2024.

RESUMO

O Infarto Agudo do Miocárdio (IAM) é uma das principais causas de morte no Brasil, especialmente entre pacientes atendidos pelo Sistema Único de Saúde (SUS). Este trabalho tem como objetivo analisar os fatores de risco associados ao desfecho (óbito ou alta) de pacientes internados por IAM no município de São Paulo, no período de 2012 a 2022, utilizando modelos de aprendizado de máquina para prever esses desfechos clínicos. Foram aplicados três modelos: Regressão Logística, Árvore de Classificação e Floresta Aleatória, com foco no desempenho preditivo e interpretabilidade. A análise revelou que, apesar do desbalanceamento dos dados, a Árvore de Classificação apresentou um desempenho mediano, mas com boa capacidade de interpretação, sendo mais adequada para contextos em que a explicação das decisões é crucial. A Floresta Aleatória, por sua vez, obteve melhores resultados preditivos, embora tenha limitado a interpretabilidade. A Regressão Logística teve desempenho inferior, mas permaneceu como um modelo útil devido à sua simplicidade. As variáveis mais significativas para os modelos foram tempo de permanência, custo total da internação e idade do paciente, cujas interações divergiram entre os modelos, indicando que fatores não lineares influenciam diretamente os desfechos analisados.

Palavras-chaves: Árvore de Classificação; Floresta Aleatória; Regressão Logística; DATASUS; Infarto Agudo do Miocárdio.

ABSTRACT

Acute Myocardial Infarction (AMI) is one of the leading causes of death in Brazil, especially among patients treated by the Unified Health System (SUS). This study aims to analyze the risk factors associated with the outcome (death or discharge) of patients hospitalized for AMI in the municipality of São Paulo between 2012 and 2022, using machine learning models to predict these clinical outcomes. Three models were applied: Logistic Regression, Classification Tree and Random Forest, with a focus on predictive performance and interpretability. The analysis revealed that, despite the unbalanced data, the Classification Tree showed average performance, but with good interpretability, being more suitable for contexts in which the explanation of decisions is crucial. Random Forest, on the other hand, obtained better predictive results, although its interpretability was limited. Logistic Regression performed less well, but remained a useful model due to its simplicity. The most significant variables for the models were length of stay, total cost of hospitalization and patient age, whose interactions differed between the models, indicating that non-linear factors directly influence the outcomes analyzed.

Keywords: Classification Tree; Random Forest; Logistic Regression; DATASUS; Acute Myocardial Infarction.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

As doenças cardiovasculares (DAC), principais responsáveis pela mortalidade no Brasil, representam uma carga significativa para o sistema de saúde, especialmente em custos com internações hospitalares (Oliveira et al., 2023). Entre essas, o infarto agudo do miocárdio (IAM) destaca-se por sua alta prevalência e taxa de mortalidade. Ainda segundo o Ministério da Saúde, o IAM é uma das principais causas de morte no país, enfatizando a importância de adotar estratégias mais eficazes de prevenção e tratamento.

O avanço tecnológico, especialmente nas áreas de análise de dados e aprendizado de máquina (AM), tem sido crucial para aprimorar o diagnóstico e prognóstico das doenças cardiovasculares. Modelos de aprendizado de máquina têm demonstrado grande potencial na predição de desfechos clínicos, facilitando a escolha de tratamentos e redução da mortalidade. "O avanço na capacidade computacional nas últimas décadas impactou especialmente o campo da detecção e predição de doenças cardiovasculares por meio da interpretação de dados [...]" (Paixão et al., 2022, p. 98), tornando essas tecnologias essenciais para identificar fatores de risco e prever desfechos clínicos.

De acordo com o Manual *Merck Sharp and Dohme* (MSD), o IAM ocorre quando há a interrupção do fluxo sanguíneo em uma artéria coronária, causando necrose no tecido cardíaco. Dependendo da extensão da lesão, pode ser classificado como transmural ou subendocárdico. O infarto subendocárdico, geralmente menos grave, ainda requer cuidados intensivos. No entanto, a gravidade do desfecho depende não apenas do tipo de infarto, mas também de fatores como idade, sexo e etnia, que influenciam a recuperação. Esses fatores tornam fundamental o uso de modelos preditivos para otimizar o tratamento e a gestão clínica de pacientes com IAM.

Este trabalho tem como objetivo a construção e avaliação de modelos de aprendizado supervisionado para prever o desfecho [1] de pacientes com infarto agudo do miocárdio subendocárdico residentes da cidade de São Paulo, utilizando dados do Sistema de Informação Hospitalar do Sistema Único de Saúde (SIHSUS). Com base nesses dados cadastrais, serão empregados três modelos amplamente utilizados: Regressão Logística (RL), Árvore de Classificação (AC) e Floresta Aleatória (FA). Esses modelos foram escolhidos por suas características complementares: a regressão logística, além de ser simples, é extremamente relevante devido à sua fácil interpretação, especialmente por meio do cálculo das *odds ratio*; a árvore de classificação, embora mais complexa computacionalmente, mantém uma alta interpretabilidade, permitindo uma visualização clara das decisões tomadas; e, por fim, a floresta aleatória, composta por múltiplas árvores de classificação, é robusta e apresenta uma excelente capacidade de generalização e alto poder preditivo.

A avaliação dos modelos será realizada utilizando métricas de desempenho oriundas da matriz de confusão e análise da curva ROC, com o objetivo de identificar qual modelo apresenta o melhor desempenho na classificação dos desfechos. Adicionalmente, será

[1] Desfecho é o resultado clínico observado após a internação, neste caso é morte ou alta hospitalar.

conduzida uma análise das variáveis mais relevantes para a predição, proporcionando *insights* sobre os principais fatores cadastrais que influenciam a recuperação e o prognóstico de pacientes com infarto agudo do miocárdio subendocárdico.

O uso de dados públicos disponibilizados pelo DATASUS apresenta uma grande vantagem, pois suas bases de dados contemplam informações representativas da população brasileira. Estudos baseados nessas informações podem contribuir para a melhoria da gestão hospitalar, o direcionamento mais eficaz de campanhas de saúde e a personalização do atendimento a pacientes com IAM, oferecendo modelos preditivos acessíveis e alinhados à realidade do SUS.

O trabalho está organizado em seis capítulos, com uma estrutura que guia o leitor de forma lógica e coesa. O primeiro capítulo apresenta uma introdução ao tema e aos métodos empregados no estudo. O segundo capítulo detalha as métricas de avaliação utilizadas para comparar os modelos. O terceiro capítulo explora a fundamentação teórica dos modelos construídos. O quarto capítulo descreve a construção da base de dados e uma análise exploratória. O quinto capítulo apresenta e discute os resultados obtidos para cada modelo. Por fim, o sexto capítulo oferece as conclusões do trabalho.

2 Métodos de Avaliação

Segundo Faceli *et al.* (2011, p. 158), "[...] a validação de qualquer nova técnica de AM proposta geralmente envolve a realização de experimentos controlados, em que se demonstre a sua efetividade na solução de diferentes problemas, representados por seus conjuntos de dados associados [...]". Nesse contexto, a experimentação controlada assegura que os modelos sejam avaliados de forma justa e confiável.

No campo do aprendizado de máquina, a avaliação de um modelo depende de práticas como a divisão dos dados em subconjuntos de treinamento e teste, para garantir que o modelo generalize bem para novos dados. Outro desafio importante é o desbalanceamento de classes, que pode afetar a performance do modelo. Técnicas específicas, como *oversampling* e *undersampling*, são comumente aplicadas para lidar com esse problema e melhorar a avaliação.

Este capítulo aborda os métodos de avaliação utilizados neste estudo. Primeiramente, será discutida a técnica de validação cruzada, seguida das soluções para o desbalanceamento de classes. Por fim, serão apresentadas as métricas de avaliação, como a matriz de confusão e a curva ROC, para medir o desempenho dos modelos de classificação.

2.1 Divisão dos dados

A divisão do conjunto de dados em subconjuntos de treinamento e teste, conhecida como *data splitting*, é uma técnica usada para avaliar o desempenho do modelo em dados novos, distintos dos usados no treinamento. Isso é importante porque, ao construir um modelo, ele pode se ajustar excessivamente aos dados de treino, resultando em superajuste *(overfitting)*. A divisão permite verificar se o modelo está superajustado aos dados de treino ou se é capaz de generalizar para dados inéditos.

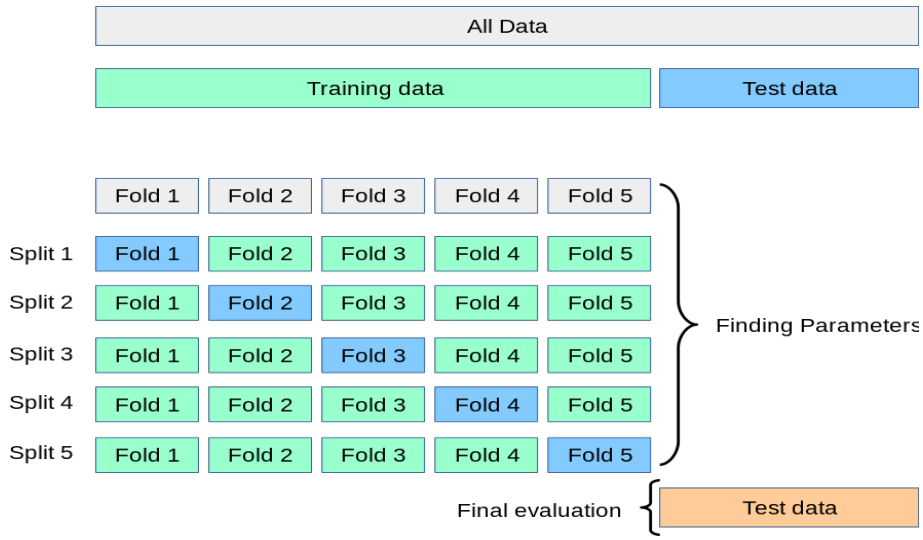
Ao realizar a divisão, a base de treinamento é maior que a base de teste, pois para treinar um modelo de forma adequada é necessário um número significativo de dados. Não existe uma regra, porém geralmente, utiliza-se uma proporção de 70% para treino e 30% para teste, especialmente em amostras medianas, ou 80% para treino e 20% para teste em amostras menores. Essas proporções podem ser ajustadas conforme o tamanho da base de dados, mas a regra é garantir que o conjunto de treino seja grande o suficiente para permitir um aprendizado robusto, enquanto o conjunto de teste deve ser representativo para avaliar a capacidade de generalização do modelo. Estas proporções em decorrência ao tamanho amostral tem como objetivo a melhor representação dos dados nas duas bases, seguindo os princípios de amostragem.

Para otimizar ainda mais a avaliação do modelo e reduzir a dependência de uma única divisão dos dados, foi desenvolvida a técnica de validação cruzada (em inglês *cross-validation*), que será apresentada no próximo tópico.

2.1.1 Validação Cruzada

A Validação Cruzada (VC) segue o princípio de dividir a base de dados em conjuntos de treinamento e teste, mas com uma metodologia mais robusta que proporciona estimativas mais confiáveis do desempenho do modelo. Define-se um número k de *folds*, que corresponde ao número de partições iguais feitas no conjunto de dados original. A cada iteração, a base de treinamento é composta por k − 1 *folds*, enquanto o *fold* restante é utilizado como conjunto de teste. Esse processo se repete até que cada *fold* tenha sido utilizado como teste uma vez. Ao final, o desempenho geral do modelo é calculado como a média das métricas obtidas em cada iteração. A Figura 1 ilustra esse processo.

Figura 1: Representação de uma validação cruzada com 5 folds



Fonte: scikit-learn **Machine Learning in Python**

Essa abordagem reduz a variabilidade nas estimativas de desempenho do modelo, já que todos os dados são utilizados tanto para treino quanto para teste, em momentos diferentes. Além disso, ela mitiga o risco de avaliações enviesadas que podem surgir de uma única divisão dos dados.

A escolha do número k depende do tamanho do conjunto de dados e do custo computacional. Um valor comumente utilizado é de k=10, porém dependendo do modelo de AM que for construído esse valor pode ser menor dado a complexidade e tempo de processamento. No contexto de classificação, a VC mantém a mesma proporção das classes em cada *fold*.

2.2 Dados desbalanceados

Em modelos de classificação, uma base de dados desbalanceada refere-se a proporções desiguais entre as classes presentes na base. Esse cenário é comum em estudos relacionados a doenças raras ou à detecção de fraudes, dado que a ocorrência da doença ou fraude (classe

minoritária) é muito menor em comparação às suas não ocorrências (classe majoritária). Construir um modelo de classificação utilizando dados desbalanceados pode resultar em um viés significativo, pois o modelo treinado tende a classificar novas observações na classe majoritária, uma vez que prioriza o acerto de elementos dessa classe.

Para contornar esse problema, existem algumas técnicas que podem ser aplicadas aos dados, como o *oversampling* e o *undersampling*. A primeira técnica consiste em replicar os dados da classe minoritária, de maneira que ela alcance o mesmo número de observações da classe majoritária ou próximo disso. Já a segunda técnica reduz o número de observações da classe majoritária para igualá-lo ao da classe minoritária ou reduzir a diferença. Ambas as abordagens resultam em uma base de dados com classes mais equilibradas, reduzindo o impacto do desbalanceamento no treinamento do modelo.

A escolha entre as duas técnicas depende de diversos fatores, como o volume dos dados disponíveis e a implementação mais adequada para o tipo de base de dados em questão. A seguir, serão apresentados os detalhes de cada uma dessas técnicas.

2.2.1 Oversampling

O aumento de observações da classe minoritária pode ser realizado por meio de uma reamostragem com reposição ou pela replicação utilizando o método SMOTE (*Synthetic Minority Oversampling Technique*), que cria instâncias sintéticas, mas plausíveis, com base nas características dos dados.

As desvantagens da reamostragem incluem a possibilidade de redundância nos dados, o que pode levar ao *overfitting*. No entanto, ele assegura a independência dos dados, o que é crucial para manter a representatividade da população. Por outro lado, o método SMOTE não garante essa independência. Por fim, o aumento das observações da classe minoritária é particularmente recomendado quando o volume de dados disponíveis para o modelo é pequeno.

2.2.2 Undersampling

O *undersampling* é uma técnica em que se reduz o número de observações da classe majoritária para equilibrar a distribuição das classes. Isso pode ser feito aleatoriamente, excluindo amostras da classe majoritária até que o número de observações se aproxime ao da classe minoritária.

Uma das vantagens do *undersampling* é que ele pode ajudar a melhorar a performance do modelo ao evitar que a classe majoritária domine o aprendizado. No entanto, essa técnica pode resultar em perda de informações importantes, já que muitas observações da classe majoritária são descartadas. Isso é particularmente problemático quando o volume total de dados é limitado. Portanto, o *undersampling* é mais recomendado quando há grande quantidade de dados disponíveis, permitindo que a redução da classe majoritária

não comprometa a capacidade de aprendizado do modelo.

2.3 Métricas avaliativas de classificação

Ao construir um modelo de classificação, diversas métricas são calculadas para avaliar seu desempenho em relação aos dados. As principais métricas são derivadas da matriz de confusão, que resume as observações corretamente e incorretamente classificadas para cada classe. Além disso, a curva ROC e sua área sob a curva (AUC) também são comumente utilizadas para medir a capacidade do modelo em distinguir entre as classes.

2.3.1 Matriz de Confusão

Este estudo utilizou dados com apenas duas classes. A seguir, será apresentada a matriz de confusão correspondente a esse contexto, juntamente com as métricas derivadas dessa matriz.

Tabela 1: Matriz de Confusão 2x2

	Predito Positivo	Predito Negativo
Real Positivo	**VP** (Verdadeiro Positivo)	**FN** (Falso Negativo)
Real Negativo	**FP** (Falso Positivo)	**VN** (Verdadeiro Negativo)

Fonte: Elaborado pelo autor

Acurácia

Mede a proporção de predições corretas.

$$Acurácia = \frac{VP + VN}{VP + FN + FP + VN}$$

Sensibilidade

Mede a proporção de positivos reais que foram corretamente identificados.

$$Sensibilidade = \frac{VP}{FN + VP}$$

Especificidade

Mede a proporção de negativos reais que o modelo identificou corretamente.

$$Especificidade = \frac{VP}{VP + VN}$$

Precisão

Mede a proporção de positivos previstos que são realmente positivos.

$$Precisão = \frac{VP}{VP + FP}$$

Kappa de Cohen

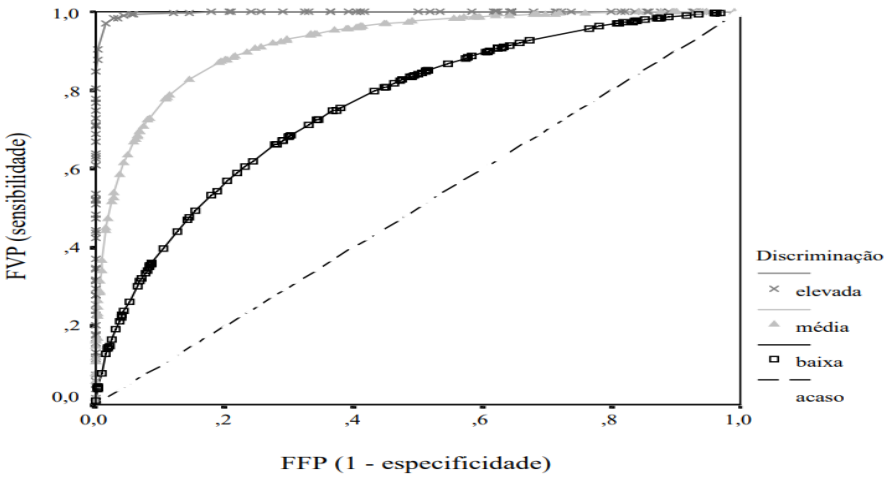
O Kappa avalia a concordância entre previsões e valores reais em dados desbalanceados, corrigindo a inflada acurácia causada pela classe majoritária, refletindo o desempenho real do modelo.

$$\kappa = \frac{2 \cdot (VP \times VN - FN \times FP)}{(VP + FP) \times (FP + VN) + (VP + FN) \times (FN + VN)}$$

2.3.2 Curva ROC

A curva *Receiver Operating Characteristic (ROC)* exibe a relação entre sensibilidade e 1 - especificidade à medida em *threshold* [2] para classificar uma observação como positiva varia. Cada ponto na curva representa um possível valor de probabilidade do modelo para prever a classe positiva, e a curva ajuda a visualizar como a alteração no ponto de corte afeta o desempenho do modelo, permitindo selecionar o melhor equilíbrio entre as classes. A Figura 2 é um exemplo de curva ROC com 3 níveis de desempenho discriminativo:

Figura 2: Curvas ROC com 3 níveis de desempenho discriminativo



Fonte: Braga(2000, p.31)

Outra medida importante é a AUC (Área sob a Curva), que avalia a capacidade do modelo em distinguir entre as classes positiva e negativa. Ela varia de 0 a 1, onde 1 indica uma classificação perfeita e 0,5 uma classificação aleatória. Quanto maior a AUC, melhor o modelo na identificação correta das classes. Para avaliar numericamente o desempenho do

[2] **threshold** é o valor de probabilidade usado para classificar uma observação como positiva ou negativa em um modelo de classificação.

modelo utilizando a curva ROC, Hosmer e Lemeshow (2013) propuseram faixas intervalares que indicam a capacidade de discriminação do modelo. As faixas estão apresentadas na Tabela 2.

Tabela 2: Regra básica para interpretação da AUC

AUC	**Discriminação**
0.5	Não tem
$0.5 < \text{AUC} < 0.7$	Fraca
$0.7 \leq AUC < 0.8$	Aceitável
$0.8 \leq AUC < 0.9$	Excelente
$AUC \geq 0.9$	Excepcional

Fonte: Tradução própria, Hosmer e Lemeshow (2013, p.177)

3 Aprendizado de máquina

Os modelos de aprendizado de máquina são algoritmos que aprendem a partir dos dados que lhes são fornecidos. Esse aprendizado pode ser dividido em três abordagens principais: **supervisionada**, não **supervisionada** e por **reforço**. Na abordagem supervisionada, o modelo é treinado com dados rotulados, ou seja, com os valores reais da variável de interesse. Em contraste, na abordagem não supervisionada, esses valores não estão disponíveis, e o modelo tenta identificar padrões ou agrupamentos subjacentes nos dados. Já o aprendizado por reforço treina o modelo com base em interações com o ambiente, buscando maximizar um objetivo de desempenho por meio de recompensas ou penalidades a cada ação realizada (Izbicki; Santos, 2020).

Dentro da abordagem supervisionada, o objetivo é aprender uma função que mapeie as entradas (dados de entrada) para a variável de saída. Especificamente, no contexto da **classificação**, existem diversos algoritmos para a construção de modelos preditivos. Dentre os mais comuns, destacam-se a **Árvore de Classificação (AC)**, baseada no algoritmo CART (Classification and Regression Tree), proposto por Breiman et al. em 1984; a **Floresta Aleatória (FA)**, que consiste em uma combinação de múltiplas árvores de decisão para melhorar a robustez e reduzir o risco de overfitting; e a **Regressão Logística (RL)**, que, apesar do nome, é amplamente usada para problemas de classificação, modelando a probabilidade de ocorrência de uma classe a partir de uma combinação linear das variáveis explicativas.

Cada um desses métodos de aprendizado de máquina possui vantagens e desvantagens, mas todos compartilham a necessidade de dividir os dados em conjuntos de **treino** e **validação**, frequentemente utilizando **validação cruzada** para garantir a generalização do modelo. Além disso, a avaliação do desempenho é feita por meio de ferramentas como as métricas de classificação derivadas da **matriz de confusão** (acurácia, precisão, sensibilidade e especificidade) e pela análise da curva ROC e da área sob a curva (AUC). O que realmente diferencia os modelos são as técnicas específicas empregadas para otimizar sua performance, como ajustes de **hiperparâmetros**.

3.1 Regressão Logística

A regressão logística é amplamente utilizada para modelar a relação entre uma variável resposta dicotômica e um conjunto de covariáveis independentes. Desde sua aceitação inicial na área epidemiológica (Hosmer; Lemeshow, 2013), o modelo tem se destacado por sua interpretabilidade, sendo amplamente utilizado em problemas de classificação, mesmo em um cenário de avanços tecnológicos.

Segundo Cramer (2002), os primeiros estudos sobre a relação logística entre variáveis ocorreram no século XIX. Desde então, inúmeros pesquisadores têm refinado essa relação, culminando no modelo que utilizamos atualmente, descrito por Cox(1958, *apud* Cramer,

2002).

Modelo de Regressão Logística

O modelo de regressão logística pertence à classe dos Modelos Lineares Generalizados (GLM), sendo adequado para variáveis dependentes dicotômicas. Cada observação, Y_i, é assumida como uma variável aleatória que segue uma distribuição Bernoulli. Essa distribuição é um caso especial da família exponencial, cuja função densidade de probabilidade pode ser escrita na forma:

$$f(y \mid \theta) = exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\}$$

Uma variável aleatória Y que tenha uma distribuição Bernoulli com probabilidade de evento igual a $\pi(x)$ tem sua função densidade de probabilidade expressa por:

$$f(x \mid \pi(x)) = P(Y \mid X) = \pi(x)^y[1 - \pi(x)]^{1-y},$$

onde $P(Y = 1 \mid X) = \pi(x)$ e $P(Y = 0 \mid X) = 1 - \pi(x)$.

A esperança de Y condicional a X é o próprio $\pi(x)$ como demonstrado abaixo:

$$E[Y|\ X] = 1 \cdot P(Y = 1|\ X) + 0 \cdot P(Y = 0|\ X) = \pi(x) + 0 \Longrightarrow E[Y|\ X] = \pi(x)$$

O modelo de regressão logística é derivado dos conceitos da regressão linear simples e múltipla, que assumem uma relação linear entre a variável resposta Y e as covariáveis X. Por exemplo, em um modelo de regressão linear múltipla, temos:

$$Y_i = X'\beta + \varepsilon_i,$$

onde β representa os coeficientes associados às covariáveis e ε_i é o erro do modelo.

No entanto, para variáveis resposta dicotômicas, como Y, que assume os valores 0 ou 1, o modelo linear direto não é apropriado. Isso ocorre porque ele pode produzir valores previstos fora do intervalo $[0, 1]$, o que não é coerente com a interpretação probabilística necessária nesse caso. Para resolver essa limitação, o modelo de regressão logística aplica a função logística no preditor linear $X'\beta$, de forma que a esperança condicional $E[Y \mid X]$ fique restrita ao intervalo $[0, 1]$. Essa aplicação é dada por:

$$E[Y \mid X] = \pi(x) = \frac{e^{X'\beta}}{1 + e^{X'\beta}},$$

Por outro lado, ao aplicarmos a função logística, a relação linear entre a variável resposta e as covariáveis é perdida. Para restabelecer essa linearidade, utiliza-se a transformação *logit*, que estabelece uma conexão entre as probabilidades $\pi(x)$ e o preditor linear $X'\beta$. Essa transformação é definida como:

$$\eta = X'\beta = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right),$$

onde η é conhecido como a função de ligação na regressão logística. Ela conecta a esperança da variável resposta, $\pi(x)$, ao preditor linear $X'\beta$, permitindo uma interpretação mais intuitiva e uma análise adequada dos parâmetros do modelo.

Estimação dos parâmetros

A estimação dos parâmetros β do modelo de regressão logística é realizada utilizando o método de máxima verossimilhança. A função de verossimilhança para o modelo é definida como:

$$L(\beta \mid X) = \prod_{i=1}^{n} f(y_i \mid x_i, \beta),$$

onde $f(y_i \mid x_i, \beta)$ é a função de probabilidade associada à variável resposta para a i-ésima observação, parametrizada por β.

A log-verossimilhança é obtida aplicando o logaritmo natural na função de verossimilhança:

$$l(\beta \mid X) = \ln\left[\prod_{i=1}^{n} f(y_i \mid x_i, \beta)\right] = \sum_{i=1}^{n} \ln\left(f(y_i \mid x_i, \beta)\right).$$

Para o modelo de regressão logística, a função de verossimilhança pode ser expressa como:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1-\pi(x_i)\right]^{1-y_i},$$

Aplicando o logaritmo natural, obtemos a função de log-verossimilhança:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} \left[y_i \ln\left(\pi(x_i)\right) + (1-y_i)\ln\left(1-\pi(x_i)\right)\right].$$

Os estimadores de máxima verossimilhança são determinados ao resolver o sistema de equações derivadas da log-verossimilhança em relação a cada parâmetro β_j, igualando essas derivadas a zero:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}\left[y_i - \pi(x_i)\right] = 0, \quad \text{para } j = 0, 1, \ldots, p,$$

onde:

- β_0 é o intercepto do modelo,
- β_j para $j = 1, \ldots, p$ são os coeficientes das covariáveis x_j,

- x_{ij} representa o valor da j-ésima covariável para a i-ésima observação.

Portanto, o sistema de equações que define os estimadores de máxima verossimilhança é:

$$\begin{cases} \displaystyle\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0, & (\text{Para } \beta_0) \\ \displaystyle\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0, & (\text{Para } \beta_j,\ j = 1, \ldots, p) \end{cases}$$

Como o modelo de regressão logística é não linear em relação aos parâmetros β, geralmente utiliza-se métodos numéricos, como o algoritmo de Newton-Raphson, para encontrar os valores que maximizam a log-verossimilhança.

Interpretação dos parâmetros

A interpretação dos parâmetros do modelo de regressão logística é feita por meio do expoente da razão de chances, também conhecida como *Odds Ratio* (OR). A *Odds Ratio* é dada pelo expoente da razão entre $\pi(x_i)$ e $1 - \pi(x_i)$, ou seja, a razão de chances para a i-ésima observação, onde x_i representa a covariável associada ao parâmetro β_i:

$$\text{OR}(\beta_i) = e^{\beta_i}.$$

Isso significa que, para cada incremento unitário em x_i, as *odds* do evento (a probabilidade de sucesso em relação ao fracasso) mudam por um fator de e^{β_i}, ou seja, se $\beta_i > 0$, o aumento de x_i faz com que as *odds* do evento aumentem; se $\beta_i < 0$, o aumento de x_i faz com que as *odds* do evento diminuam, se a *odds* for 1 a covariável não é significativa no modelo, quando construído o intervalo de confiança das OR, se caso o valor 1 estiver no intervalo a covariável não é signficativa. A fórmula do intervalo de confiança (IC) para a odds ratio (OR) é dada por:

$$IC(OR(\beta)) = \exp\left(\hat{\beta} \pm z_{1-\alpha/2} \cdot \text{SE}(\hat{\beta})\right),$$

onde $\hat{\beta}$ é o estimador do coeficiente, $SE(\hat{\beta})$ é o erro padrão de $\hat{\beta}$ e $z_{1-\alpha/2}$ é o valor crítico da distribuição normal padrão.

3.1.1 Significância dos coeficientes

Assim como em um modelo de regressão linear, as covariáveis no modelo precisam apresentar significância estatística, ou seja, devem explicar a variação dos dados. Caso contrário, não há motivo para mantê-las no modelo. Na regressão logística, essa significância é avaliada por meio do teste de razão de verossimilhança e do teste de *Wald*.

Teste da Razão de Verossimilhança

O teste da razão de verossimilhança é um teste global que verifica se ao menos um dos parâmetros estimados é significativo no modelo. Ele compara a log-verossimilhança entre dois modelos: um modelo nulo (sem as covariáveis) e um modelo alternativo (com as covariáveis). A estatística do teste é dada pela diferença nas log-verossimilhanças dos modelos, multiplicada por -2. As hipóteses e a fórmula da estatística do teste são as seguintes:

$$\begin{cases} H_0 : \beta = 0 \quad \text{(nenhuma covariável é significativa)} \\ H_a : \beta \neq 0 \quad \text{(pelo menos uma covariável é significativa)} \end{cases}$$

A estatística de razão de verossimilhança é dada por:

$$D = -2 \ln \left[\frac{\text{verossimilhança do modelo nulo}}{\text{verossimilhança do modelo alternativo}} \right]$$

Essa estatística segue uma distribuição qui-quadrado (χ^2) com graus de liberdade (g.l.) iguais ao número de parâmetros estimados (p), excluindo o intercepto (g.l.$= p - 1$). O teste verifica se a adição de covariáveis melhora significativamente o ajuste do modelo em comparação com um modelo sem as covariáveis. Se o valor de D for suficientemente grande, rejeita-se a hipótese nula, indicando que ao menos uma covariável é significativa.

Teste de Wald

O teste de Wald é um teste individual que verifica a significância de um parâmetro estimado em um modelo. Ele é baseado na razão entre a estimativa do parâmetro e o erro padrão da estimativa. Esse teste é útil quando se deseja testar a significância de um único parâmetro. As hipóteses e a estatística do teste são as seguintes:

$$\begin{cases} H_0 : \beta_j = 0 \quad \text{(o parâmetro não é significativo)} \\ H_a : \beta_j \neq 0 \quad \text{(o parâmetro é significativo)} \end{cases}$$

A estatística de Wald é dada por:

$$W = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} \sim N(0, 1)$$

O valor W segue uma distribuição normal padrão $N(0, 1)$, e o quadrado de W, ou seja, W^2, segue uma distribuição qui-quadrado (χ^2) com 1 grau de liberdade. Se o valor de W^2 for maior que o valor crítico da distribuição qui-quadrado para o nível de significância escolhido, rejeita-se a hipótese nula, indicando que o parâmetro β_j é significativo no modelo.

Ambos os testes, o teste da razão de verossimilhança e o teste de Wald, fornecem informações importantes sobre a qualidade do ajuste do modelo e a significância dos

parâmetros. O teste da razão de verossimilhança é geralmente utilizado para avaliar a importância global do modelo, enquanto o teste de Wald é mais adequado para testar a significância de parâmetros individuais. Ambos devem ser considerados ao validar um modelo de regressão logística.

3.1.2 Otimização do modelo de regressão

A otimização de um modelo de regressão envolve a escolha do conjunto de variáveis que melhor se ajustam aos dados, e para isso, pode-se utilizar técnicas como o método *stepwise*, que funciona de maneira semelhante à poda de uma árvore. Esse método possui três abordagens: na abordagem *backward*, começa-se com um modelo completo e remove-se variáveis uma a uma, avaliando a métrica (geralmente o AIC [3]) até encontrar o modelo ideal; na abordagem *forward*, inicia-se com um modelo simples, adicionando variáveis progressivamente até chegar ao modelo ótimo; e na abordagem *backward-forward*, combina-se as duas anteriores, alternando entre adicionar e remover variáveis para encontrar o equilíbrio entre simplicidade e completude.

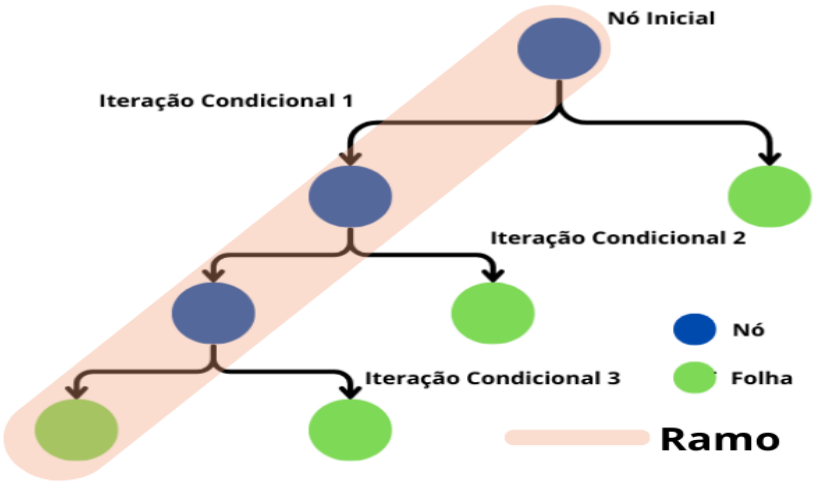
Uma abordagem alternativa para otimizar um modelo de regressão é por meio da regularização, que pode ser realizada de três formas principais: L1 (*LASSO*), L2 (*Ridge*) ou uma combinação de ambos, conhecida como *ElasticNet*. Esses métodos são comumente chamados de técnicas de tuning, pois aplicam restrições à função de estimação dos coeficientes do modelo. A regularização LASSO tende a zerar os coeficientes de variáveis que não contribuem de forma significativa para a explicação dos dados. Por outro lado, a regularização Ridge não zera os coeficientes, mas reduz substancialmente os valores daqueles que possuem menor impacto. O *ElasticNet* combina os dois métodos, buscando aproveitar as vantagens de ambos. Detalhes mais aprofundados sobre essas técnicas podem ser encontrados em (Hastie; Tibshirani e Friedman, 2008).

[3] O AIC (*Akaike Information Criterion*) é uma métrica usada para comparar modelos estatísticos, balanceando ajustes e simplicidade, penalizando a complexidade excessiva.

3.2 Árvores de classificação

Em seu livro (Silva; Peres e Boscarioli, 2016) afirmam: "[...] No contexto da resolução da tarefa de classificação, uma árvore de decisão representa o modelo capaz de guiar a tomada de decisão sobre a determinação da classe à qual um exemplar pertence". A árvore de classificação é um modelo de classificação hierárquico composto por nós, folhas e ramos, conforme a figura 3. Os nós das árvores representam um processo de iteração condicionada das covariáveis, resultando em partições que levam a outros nós e finalizam em uma folha, que determina a classe à qual o elemento pertence segundo o modelo, o percurso que cada elemento fez até chegar na folha é chamado de ramo.

Figura 3: Representação de uma árvore de classificação



Fonte: Elaborado pelo autor

As divisões são realizadas em uma covariável por vez. Em cada nó que os dados atravessam, eles são divididos em subconjuntos distintos. Cada partição da árvore forma subconjuntos distintos e disjuntos, representados por R_k, sendo k o número indicativo da partição. Essa divisão em subconjuntos é o que permite a classificação da variável resposta de cada elemento. No caso de uma regressão, a classe escolhida pode ser dada como a média dos valores da variável resposta dos elementos da folha; porém, em uma árvore de classificação (AC), é comum usar a moda para verificar a classe mais frequente da folha.

$$g(x) = moda\,\{y_i | x_i \in R_k\}$$

As partições dos dados em cada nó têm o propósito de separar os indivíduos em conjuntos homogêneos. Para isso, são utilizados certos critérios que avaliam qual iteração vai melhor separar os indivíduos. Na árvore de regressão, pode ser utilizado o critério do erro quadrático médio, enquanto na classificação podem ser utilizados os critérios de **Índice Gini** ou de **entropia**.

Índice de Gini

O índice de Gini, proposta por Breiman et. al (1984) é uma métrica amplamente utilizada para medir a impureza de um nó em árvores de classificação. Ele avalia a probabilidade de um elemento ser classificado incorretamente ao ser escolhido aleatoriamente, considerando a distribuição das classes no nó. Quanto menor o índice de Gini, maior a homogeneidade do nó, o que indica maior pureza. Assim, ele é utilizado como critério para determinar as divisões que tornam os subconjuntos mais homogêneos.

Em um contexto onde o modelo visa classificar a morte ou sobrevivência de um paciente, ou seja apenas duas classes, o índice de Gini é definido matematicamente como:

$$Gini = 1 - \sum_{i=1}^{2} p_i^2,$$

onde p_i é a proporção de elementos pertencentes à classe i no nó.

O índice de Gini varia entre 0 e 1. Quando o índice de Gini é igual a 0, o nó é considerado puro, ou seja, todos os elementos pertencem à mesma classe. Já quando o índice de Gini é igual a 1, o nó é considerado totalmente impuro, indicando que metade das observações pertence a uma classe e a outra metade à outra classe.

Entropia e Ganho de Informação

A entropia é um conceito amplo que aparece em diferentes áreas do conhecimento. Na *Teoria da Informação*, a entropia é usada como uma medida de incerteza ou desordem de um conjunto de dados. No contexto de uma Árvore de Classificação, a entropia mede o grau de impureza dos nós, ou seja, avalia a eficácia de uma divisão.

Quanto maior a entropia, maior a incerteza sobre as classes no nó, o que indica maior impureza. Por outro lado, uma entropia menor reflete maior certeza de que os elementos pertencem a uma única classe, indicando maior pureza do nó. Assim, o objetivo do algoritmo ao dividir os nós é minimizar a entropia, gerando grupos mais homogêneos.

No contexto de classificação binária, como no caso de prever a morte ou sobrevivência de um paciente, a entropia de um nó pode ser simplificada para duas classes. A fórmula da entropia para um nó com duas classes é dada por:

$$H(S) = -p_1 \log_2(p_1) - p_2 \log_2(p_2),$$

onde p_i é a proporção de elementos pertencentes à classe i no nó.

A entropia varia de 0 a 1. Quando as classes estão igualmente distribuídas ($p_1 = p_2 = 0.5$), a entropia é máxima ou seja igual a 1, indicando alta incerteza sobre a classificação. Por outro lado, quando todas as observações no nó pertencem a uma única classe ($p_1 = 1$ e $p_2 = 0$ ou $p_1 = 0$ e $p_2 = 1$), a entropia é mínima, indicando um nó puro.

O **Ganho de Informação** segue o mesmo raciocínio da entropia: ele representa a diferença na entropia antes e depois de uma divisão no nó, indicando o quanto de incerteza

foi reduzida. Em outras palavras, o ganho de informação mede a quantidade de informação que foi adquirida ao realizar a divisão. Assim, para maximizar o ganho de informação, é escolhida a divisão que resulta na menor entropia, ou seja, aquela que gera nós mais homogêneos, com maior pureza nas classes.

3.2.1 Exemplo de Árvore de classificação

Considere os dados apresentados na Tabela 3, que contém informações de 10 pacientes. Cada paciente é classificado de acordo com a presença ou ausência de uma doença (variável **Classe**), além de informações adicionais sobre o **Sexo** e se o paciente é **Fumante**. Esses dados serão utilizados para calcular o Índice de Gini e a Entropia, com o objetivo de determinar qual variável inicial deve ser utilizada para dividir o nó na construção da árvore de classificação.

Tabela 3: Dados do exemplo de demonstração

Paciente	**Classe**	**Sexo**	**Fumante**
1	Positivo	Masculino	Sim
2	Negativo	Feminino	Sim
3	Negativo	Feminino	Sim
4	Positivo	Masculino	Não
5	Positivo	Masculino	Não
6	Negativo	Feminino	Sim
7	Positivo	Masculino	Não
8	Negativo	Feminino	Sim
9	Negativo	Feminino	Sim
10	Negativo	Feminino	Não

Fonte: Elaborado pelo autor

A tabela permite calcular a pureza dos nós considerando as duas variáveis (**Sexo** e **Fumante**).

Cálculo do Índice de Gini e Entropia por Sexo

Distribuição por Sexo:

- Masculino (*Positivo:* 4, *Negativo:* 0): $p_{Positivo} = 1.0,\ p_{Negativo} = 0.0$
- Feminino (*Positivo:* 0, *Negativo:* 6): $p_{Positivo} = 0.0,\ p_{Negativo} = 1.0$

Índice de Gini (Sexo):

Para Masculino:

$$Gini_M = 1 - (1.0^2 + 0.0^2) = 0$$

Para Feminino:

$$Gini_{Fem} = 1 - (0.0^2 + 1.0^2) = 0$$

Gini total:

$$Gini = \frac{4}{10} \cdot 0 + \frac{6}{10} \cdot 0 = 0$$

Entropia (Sexo):
Para Masculino:

$$Entropia_M = -(1.0 \cdot \log_2(1.0) + 0.0 \cdot \log_2(0.0)) = 0$$

Para Feminino:

$$Entropia_{Fem} = -(0.0 \cdot \log_2(0.0) + 1.0 \cdot \log_2(1.0)) = 0$$

Entropia total:

$$Entropia = \frac{4}{10} \cdot 0 + \frac{6}{10} \cdot 0 = 0$$

Cálculo do Índice de Gini e Entropia por Fumante

Distribuição por Fumante:

- Fumante (*Positivo:* 1, *Negativo:* 5): $p_{Positivo} = 0.1667,\ p_{Negativo} = 0.8333$
- Não Fumante (*Positivo:* 3, *Negativo:* 1): $p_{Positivo} = 0.75,\ p_{Negativo} = 0.25$

Índice de Gini (Fumante):
Para Fumante:

$$Gini_{Fum} = 1 - (0.1667^2 + 0.8333^2) = 1 - (0.028 + 0.694) = 0.278$$

Para Não Fumante:

$$Gini_{NaoFum} = 1 - (0.75^2 + 0.25^2) = 1 - (0.5625 + 0.0625) = 0.375$$

Gini total:

$$Gini = \frac{6}{10} \cdot 0.278 + \frac{4}{10} \cdot 0.375 = 0.3168$$

Entropia (Fumante):
Para Fumante:

$$Entropia_{Fum} = -(0.1667 \cdot \log_2(0.1667) + 0.8333 \cdot \log_2(0.8333)) = 0.65$$

Para Não Fumante:

$$Entropia_{NaoFum} = -(0.75 \cdot \log_2(0.75) + 0.25 \cdot \log_2(0.25)) = -(-0.3113 - 0.5) = 0.8113$$

Entropia total:

$$Entropia = \frac{6}{10} \cdot 0.65 + \frac{4}{10} \cdot 0.8113 = 0.71452$$

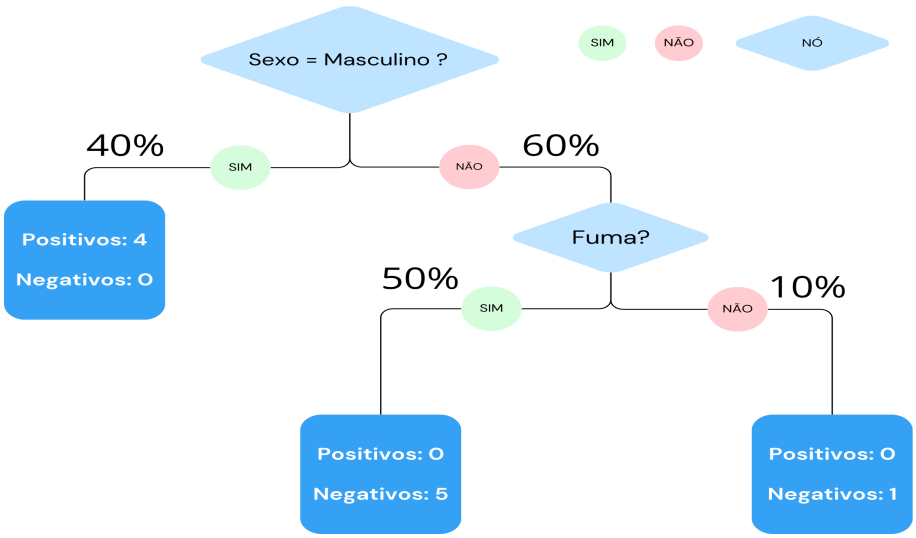
Escolha do Nó Inicial

Com base nos cálculos, temos:

- **Índice de Gini:** Sexo (0.0) é mais puro que Fumante (0.3168).
- **Entropia:** Sexo (0.0) é mais puro que Fumante (0.71452).

Portanto, a variável **Sexo** deve ser utilizada como nó inicial, tanto pelo índice de Gini quanto pela Entropia, dado que sexo é a variável que melhor divide os dados. A figura 4 demonstra como fica a árvore de classificação construída.

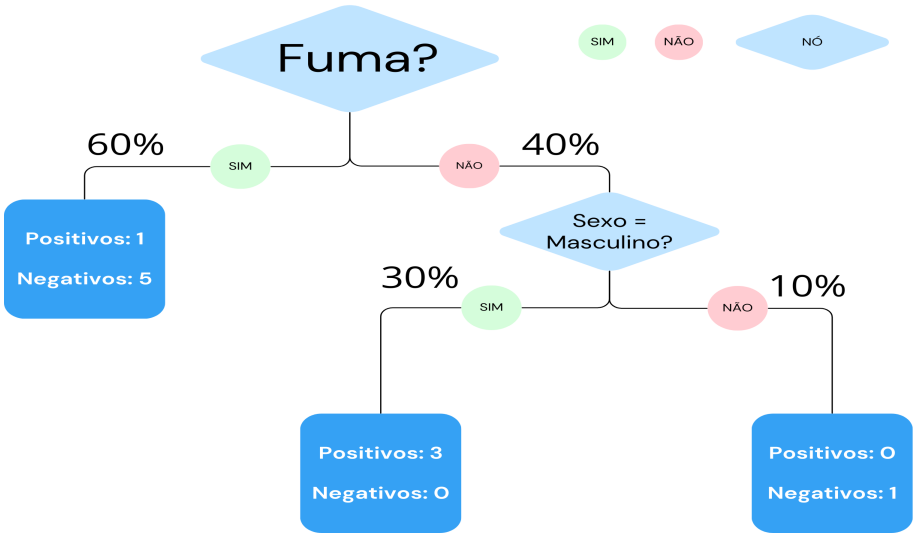
Figura 4: Árvore de classificação do exemplo



Fonte: Elaborado pelo autor

Por se tratar de um exemplo, a árvore gerada é bastante simples. No entanto, é possível observar que todas as folhas contêm dados homogêneos, o que indica que a divisão com base no índice de Gini ou na entropia consegue discriminar bem as classes. A Figura 5 mostra a árvore gerada quando os índices de Gini ou de entropia não são utilizados.

Figura 5: Árvore de classificação do exemplo com nó inicial errado



Fonte: Elaborado pelo autor

Ao analisar as folhas, observa-se que uma delas não é homogênea, o que comprova os princípios das técnicas de divisão dos nós.

3.2.2 Otimização da árvore por poda

Uma das vantagens das Árvores de Classificação é o seu poder interpretativo e bom desempenho na classificação. A interpretação do modelo é feita por meio dos ramos da árvore. No entanto, à medida que a árvore cresce, a interpretação torna-se mais complexa e, em níveis mais profundos, as divisões podem se tornar mais específicas e menos intuitivas. Dessa forma, é mais eficaz reduzir o tamanho da árvore, o que também melhora a generalização do modelo para novos dados, evitando o superajuste. Esse processo é conhecido como poda (ou *pruning*, em inglês), e existem dois tipos principais: **pré-poda** e **pós-poda**.

Pré-poda

De acordo com Silva, Peres e Boscarioli (2011, p. 99), "A pré-poda conta com regras de parada que previnem a construção daqueles ramos que não parecem melhorar a precisão preditiva da árvore. [...]". Basicamente, a poda é realizada limitando os hiperparâmetros da árvore, como a profundidade máxima, o número mínimo de observações por nó e o índice de impureza mínimo. Essas restrições evitam que a árvore se torne excessivamente complexa, o que pode reduzir o risco de overfitting e melhorar a interpretabilidade do modelo. No entanto, a pré-poda não é tão amplamente utilizada, pois diversos pesquisadores chegaram a um consenso de que limitar os hiperparâmetros antes da construção da árvore pode acarretar uma grande perda de informação de forma precoce, sem considerar adequadamente as interações entre as covariáveis. Isso ocorre porque, ao interromper o crescimento da

árvore antes de sua construção completa, pode-se impedir que o modelo capture padrões e interações importantes nos dados, prejudicando sua capacidade de generalização.

Pós-poda

A pós-poda é uma técnica que começa com a construção de uma árvore de decisão completa, sem restrições iniciais, e, em seguida, realiza a poda de seus ramos com base em critérios específicos, como o erro de classificação ou o custo de complexidade. Este último é amplamente utilizado e envolve a geração de subárvores progressivamente menores, representadas por $|T|$. Para cada subárvore, calcula-se a taxa de erro $R(T)$ ajustada pelo tamanho da árvore. O custo de complexidade (em inglês, *Complexity Price* ou CP) é definido pela fórmula:

$$R_\alpha(T) = R(T) + \alpha|T|,$$

onde α é um parâmetro que pondera a importância do tamanho da árvore em relação à taxa de erro.

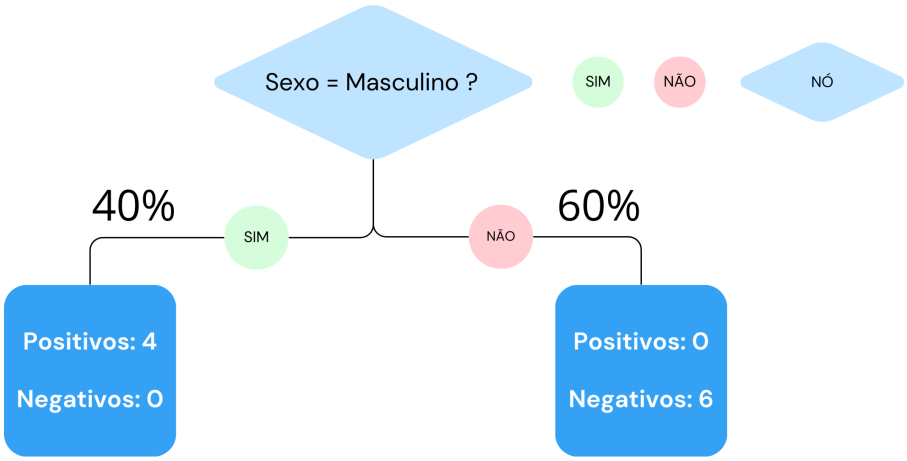
O parâmetro α atua como um fator de penalização, controlando o equilíbrio entre a complexidade e a precisão da árvore. Esse equilíbrio pode ser descrito como:

$$\begin{cases} \text{Para } \alpha \text{ pequeno, a árvore será maior (mais complexa);} \\ \text{Para } \alpha \text{ grande, a árvore será menor (menos complexa).} \end{cases}$$

O objetivo final é determinar o valor de α que minimiza $R_\alpha(T)$, identificando assim a subárvore que alcança o melhor equilíbrio entre desempenho e simplicidade.

Dessa forma, ao voltarmos ao exemplo anterior, mostrado na Figura 4, é possível perceber que a árvore poderia ter sido podada já no nó raiz, uma vez que, com uma única divisão, ele geraria duas folhas homogêneas e um modelo menos complexo. A Figura 6 ilustra isso.

Figura 6: Árvore de classificação do exemplo podada



Fonte: Elaborado pelo autor

3.2.3 Otimização do modelo por *tuning*

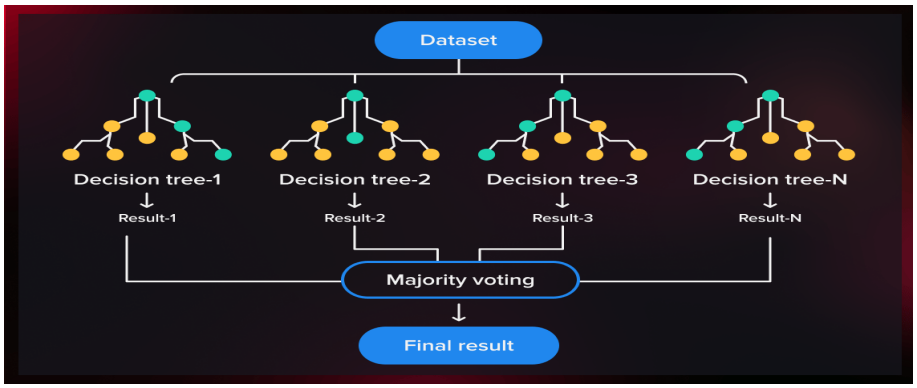
Uma forma de otimizar o modelo de Árvore de classificação é utilizando *tuning* nos hiperparâmetros, ou seja, ajustando os valores ótimos para esses parâmetros. Esse processo é semelhante aos conceitos de pré-poda e pós-poda, pois envolve ajustes em parâmetros como a profundidade da árvore, o número mínimo de observações por nó e o custo de complexidade.

O processo de *tuning* define diferentes valores para cada hiperparâmetro do modelo e avalia o desempenho dos modelos gerados com esses parâmetros. Para cada combinação, são calculadas as principais métricas de desempenho, como acurácia, precisão, sensibilidade, entre outras. O valor ótimo do hiperparâmetro é aquele que resulta no melhor desempenho do modelo, ou seja, o modelo que maximiza as métricas escolhidas. Esse processo ajuda a identificar a configuração mais eficiente e robusta para o modelo, melhorando sua capacidade preditiva.

3.3 Florestas aleatórias

Segundo Morettin e Singer (2021, p. 327), "De um modo geral, árvores de decisão produzem resultados com grande variância [...]". Esse comportamento ocorre porque uma única árvore tende a se ajustar muito aos dados de treinamento, tornando-a sensível a pequenas variações na base de dados. Para mitigar esse problema, foi desenvolvido o modelo de Florestas Aleatórias (FA), que combina múltiplas árvores de decisão em um esquema de aprendizado em conjunto, utilizando uma técnica chamada *bagging*.

Figura 7: Representação de uma floresta aleatória

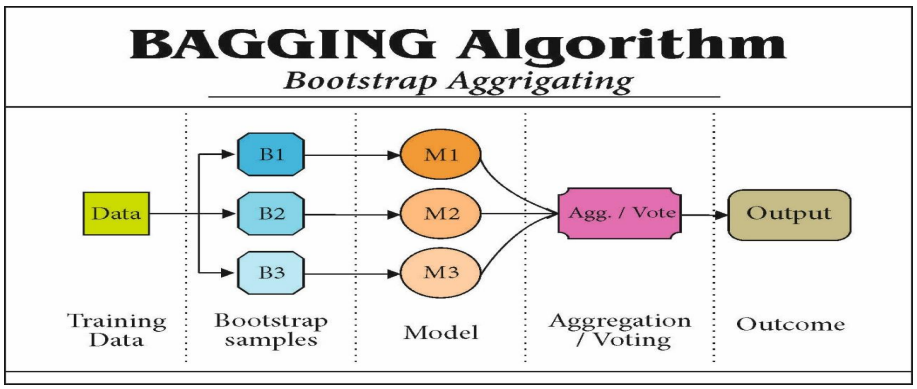


Fonte: Serokell, 2022

O *bagging*, abreviação de *Bootstrap Aggregation*, baseia-se na ideia de treinar cada árvore em uma amostra *bootstrap* (amostras aleatórias com reposição) da base de treinamento original, a figura 8 apresenta o algoritmo do método. Cada árvore é ajustada de forma independente, gerando uma sequência de modelos especializados. Além disso, para

aumentar a diversidade entre as árvores, as florestas aleatórias introduzem aleatoriedade adicional ao selecionar, em cada nó, um subconjunto aleatório de n covariáveis (com n<p) das p variáveis disponíveis. A melhor divisão é então escolhida com base nesse subconjunto reduzido, evitando o domínio de variáveis preditoras muito influentes.

Figura 8: Algoritmo do método *bagging*



Fonte: Analytics India Magazine, 2024

O resultado final da Floresta Aleatória é determinado pela agregação dos resultados individuais de todas as árvores, como é demonstrado na figura 7. Para problemas de classificação, essa agregação é feita pela moda das classes previstas pelas árvores (ou seja, a classe mais votada), enquanto em problemas de regressão, utiliza-se a média das previsões. Essa combinação reduz a variância do modelo e melhora sua capacidade de generalização, já a aleatorização das variáveis nos nós reduz a correlação entre as árvores.

Embora as Florestas Aleatórias contornem os principais defeitos das árvores de classificação, o modelo apresenta algumas limitações, sendo as principais o alto custo computacional e a perda de interpretabilidade em troca de maior poder preditivo e robustez.

3.4 Interpretação do modelo

As florestas aleatórias oferecem uma forma de interpretação mínima por meio das importâncias atribuídas às variáveis. Essas importâncias são calculadas considerando todas as árvores construídas e avaliando como cada variável contribui para reduzir o erro de predição ou a impureza nos nós. Duas métricas principais são usadas para essa interpretação: a redução média da precisão (*Mean Decrease in Accuracy*), que avalia o impacto da variável no desempenho global do modelo, e a redução média da impureza (*Mean Decrease in Gini*), que mede o ganho de pureza proporcionado pelas divisões associadas à variável. Essas métricas permitem identificar quais variáveis são mais relevantes para as decisões do modelo, mesmo que a floresta aleatória seja essencialmente uma técnica de caixa-preta [4].

[4] Técnicas de caixa-preta, ou *black box*, referem-se a modelos cujo funcionamento interno é complexo e pouco interpretável, mas que geram resultados altamente precisos

3.4.1 Otimização do modelo por *tuning*

Assim como as Árvores de Classificação, as Florestas Aleatórias também utilizam estratégias de otimização do modelo. Isso é feito principalmente por meio da definição e ajuste de seus hiperparâmetros, com o objetivo de reduzir o custo computacional sem comprometer significativamente o desempenho preditivo. O custo computacional é frequentemente medido pelo tempo necessário para a construção do modelo. Os principais hiperparâmetros envolvidos na otimização do modelo são:

Número de árvores na floresta: Quanto maior o número de árvores no modelo, melhor será a precisão, mas isso também eleva o custo computacional. Ajustar o número de árvores é importante para equilibrar o custo computacional com o poder preditivo do modelo. Uma quantidade excessiva de árvores pode resultar em um custo elevado sem ganhos significativos em termos de precisão, especialmente após um número de árvores suficiente para garantir boa generalização.

Profundidade máxima das árvores: A profundidade máxima limita o número de divisões em cada árvore. Árvores muito profundas podem capturar ruídos e peculiaridades do conjunto de treinamento, causando *overfitting*. Ajustar a profundidade é essencial para balancear a complexidade do modelo e sua capacidade de generalização. Uma profundidade muito baixa pode resultar em modelos muito simples, enquanto profundidades excessivas podem levar a modelos excessivamente complexos e propensos ao *overfitting*.

Número de variáveis a serem usadas em cada nó: Um número elevado de variáveis escolhidas aleatoriamente em cada nó aumenta o custo computacional, mas pode contribuir para uma maior diversidade entre as árvores, ajudando a evitar o *overfitting*. Ajustar esse parâmetro é importante para equilibrar a diversidade das árvores na floresta e o custo computacional, garantindo que o modelo tenha boa capacidade preditiva sem ser excessivamente complexo.

Número máximo de nós por árvore: Uma árvore com muitos nós aumenta a complexidade da floresta e o risco de *overfitting*. No entanto, um número muito baixo de nós pode resultar em árvores muito rasas, com baixo poder preditivo. Ajustar o número máximo de nós ajuda a controlar a complexidade do modelo e a balancear o ajuste ao conjunto de treinamento e a capacidade de generalização do modelo.

A partir dos hiperparâmetros apresentados acima, realiza-se o processo de *tuning* para encontrar a melhor configuração para a floresta aleatória. O objetivo do *tuning* é ajustar esses hiperparâmetros de forma a otimizar o desempenho do modelo, equilibrando o poder preditivo e o custo computacional.

4 Tratamento dos dados

Neste trabalho, os modelos foram construídos a partir de dados do Sistema de Informações Hospitalares do Sistema Único de Saúde (SIHSUS), disponibilizados pelo Departamento de Informática do SUS (DATASUS). A base abrange registros do estado de São Paulo entre 2012 a 2022.

No presente estudo, o desfecho analisado foi o infarto agudo do miocárdio subendocárdico (CID-10: I21.4), uma condição de alta relevância clínica e epidemiológica. Foram incluídos na base de dados apenas pacientes residentes no município de São Paulo com registro da idade em anos, para garantir consistência geográfica e populacional na análise.

A seleção das covariáveis envolveu etapas de filtragem para remover variáveis sem contribuição significativa ao estudo. Inicialmente, foram excluídas variáveis com mais de 45% de valores ausentes, eliminando 18 variáveis. Em seguida, foram removidas aquelas com baixa variabilidade, definidas como variáveis com mais de 90% das observações concentradas no mesmo valor, resultando na exclusão de outras 53 variáveis.

Após essas etapas, restaram 42 variáveis, das quais 30 eram informações administrativas ou identificadores sem relevância direta para os modelos. Essas também foram descartadas, deixando uma base final com 12 variáveis potencialmente úteis.

A Figura 9 ilustra a redução progressiva no número de variáveis ao longo do tratamento. Todas as etapas de preparação e limpeza de dados foram realizadas na linguagem *R*, utilizando os pacotes do *tidyverse*.

Figura 9: Filtragem das variáveis originais

Fonte: Elaborado pelo autor

Entre as 12 variáveis remanescentes, quatro estavam relacionadas aos custos das internações: Custo de Serviços Profissionais, Custo de Serviços Hospitalares, Custo Total e Custo de Unidade de Tratamento Intensivo (UTI). Devido à alta correlação entre essas variáveis, optou-se por manter apenas o Custo Total, que representa a soma das demais, e o Custo de UTI, considerando sua relevância e valores significativamente mais elevados.

A Tabela 4 apresenta as variáveis finais selecionadas para o modelo, já incluindo o

tratamento das variáveis qualitativas por meio de *dummies*. A variável "Morte"foi definida como a variável de interesse para o estudo.

Tabela 4: Variáveis selecionadas e tratadas

Variável	**Descrição**
Idade	Idade do paciente, em anos.
Sexo	Sexo do paciente: 1 - Masculino, 0 - Feminino.
Permanência	Quantidade de dias que o paciente permaneceu internado.
UTI	Internação em UTI: 1 - Sim, 0 - Não.
Dias na UTI	Quantidade de dias que o paciente permaneceu internado na UTI.
Complexidade	Procedimento de alta ou média complexidade: 1 - Alta, 0 - Média.
Urgência	Internação por urgência: 1 - Sim, 0 - Não.
Custo Total	Valor total gasto na internação de um paciente.
Custo UTI	Valor gasto na internação de um paciente em uma UTI.
Morte	**Desfecho do paciente: 1 - Óbito, 0 - Alta.**

Fonte: Elaborado pelo autor.

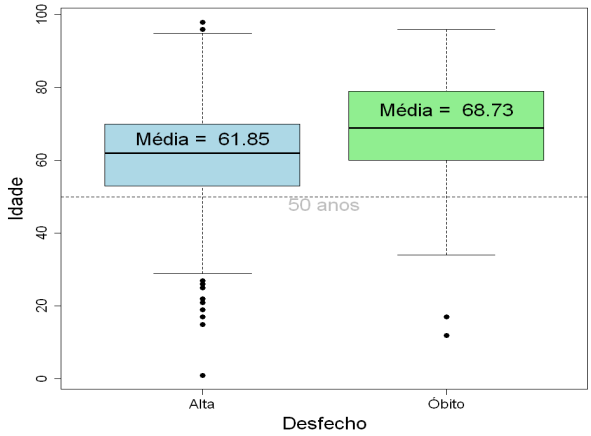
4.1 Análise Exploratória dos Dados

A Análise Exploratória de Dados (AED) busca compreender a estrutura e o comportamento dos dados antes da aplicação de modelos estatísticos ou de aprendizado de máquina. Essa etapa inclui a análise de distribuições, identificação de *outliers* e a visualização dos dados visando revelar padrões e relações importantes. Além disso, a AED é crucial para verificar os pressupostos dos modelos, garantindo ajustes adequados e resultados confiáveis.

Nos modelos de classificação binária desenvolvidos neste trabalho, a AED foi realizada separadamente para cada classe, permitindo comparar o comportamento das variáveis entre elas. Essa análise ajuda a identificar padrões importantes, como variáveis capazes de distinguir bem as classes ou casos de sobreposição que podem impactar a performance do modelo.

A Figura 10 revela que pacientes que receberam alta possuem uma idade média inferior à dos pacientes que faleceram, indicando que a idade pode ser um fator de risco importante para o IAM. Também se observa que **a maioria das internações envolve pacientes com mais de 50 anos**, destacando a relevância dessa faixa etária na análise.

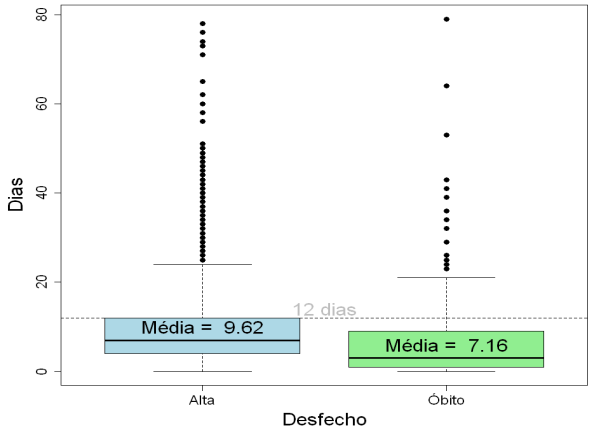
Figura 10: Boxplot das idades dos pacientes por desfecho



Fonte: Elaborado pelo autor

Observando a Figura 11, nota-se que **a maioria dos pacientes permanecem internado por até 12 dias**. Entre os que recebem alta, o tempo médio de internação é dois dias e meio maior do que o dos pacientes que falecem. No entanto, sob influência dos *outliers* a média se torna menos confiável, pois é sensível a valores extremos. Por isso, utiliza-se a mediana: para pacientes com alta, a mediana é de 7 dias, enquanto para os que faleceram é de 3 dias, sugerindo que os pacientes que falecem tendem a ter internações significativamente mais curtas.

Figura 11: Boxplot da quantidade de dias internados por desfecho



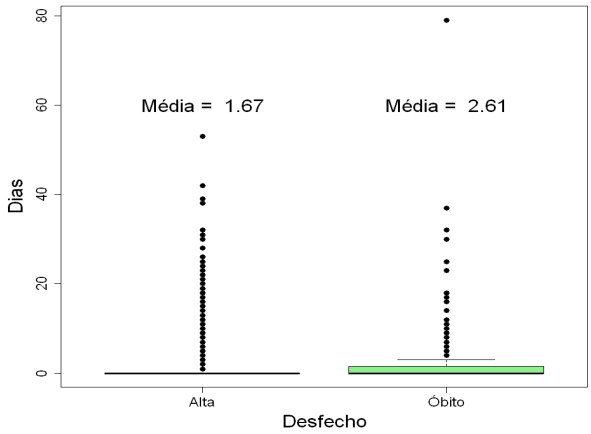
Fonte: Elaborado pelo autor

O gráfico da Figura 12 evidencia a presença de *outliers* em ambas as classes analisadas. Observa-se que, em média, **os pacientes que falecem permanecem internados quase um dia a mais do que os que recebem alta**. No entanto, a média e a mediana não é uma métrica confiável nesse caso devido à alta frequência de valores zero gerando uma **mediana nula** e dos *outliers*, que distorcem a interpretação dos dados. Para avaliar a diferença entre as classes, foi aplicado o teste de *Mann-Whitney* [5], que resultou em

[5] **Teste de Mann-Whitney** é um teste não paramétrico que compara duas distribuições para determinar se elas diferem significativamente em termos de posição central.

um p-valor de 0,144. Esse resultado não permite rejeitar a hipótese nula de ausência de diferença significativa entre as distribuições das duas classes.

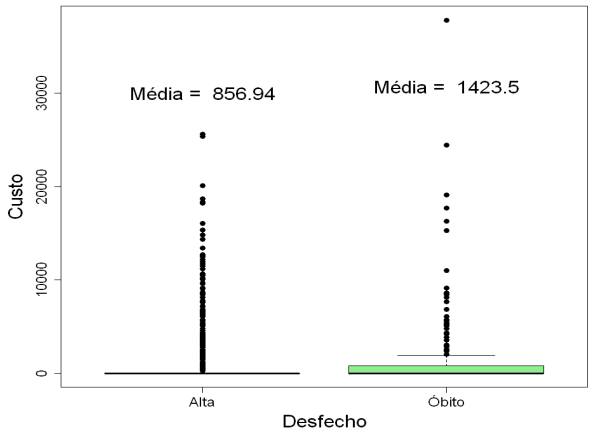
Figura 12: Boxplot da quantidade dias na UTI por desfecho



Fonte: Elaborado pelo autor

O gráfico da Figura 13 tem a mesma interpretação da Figura 12, pois a maioria dos pacientes não foi internada na UTI, e, portanto, não gerou custos. No entanto, o teste de *Mann-Whitney* foi realizado, já que os custos de UTI podem variar. O teste retornou um p-valor de 0.1362, não rejeitando a hipótese nula de que não há diferença significativa entre as classes.

Figura 13: Boxplot dos custos de UTI por desfecho

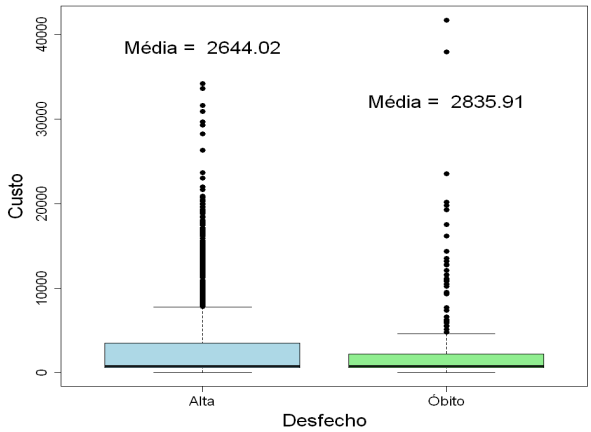


Fonte: Elaborado pelo autor

Ao observar a Figura 14, percebe-se que a média dos custos dos pacientes que foram a óbito é maior do que a dos pacientes que receberam alta, influenciada pela presença de *outliers*. No entanto, ao analisar as medianas, verifica-se que o custo dos pacientes que receberam alta é superior ao daqueles que faleceram, com valores de 808,8 e 778,3, respectivamente. Isso pode indicar que os pacientes que recebem alta estão associados

a tratamentos mais caros ou que o tempo adicional de internação, em média dois dias, contribui para um aumento considerável no custo total.

Figura 14: Boxplot dos custos totais por desfecho

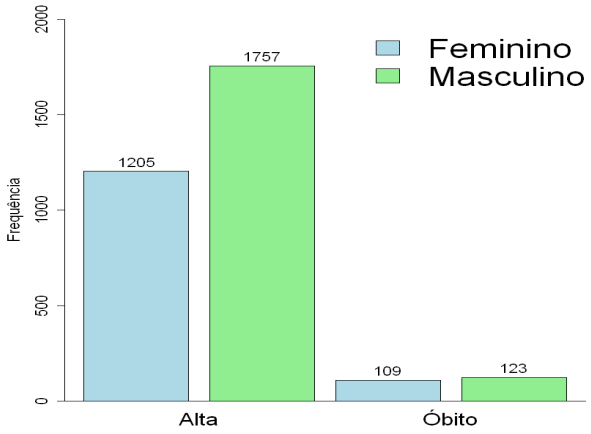


Fonte: Elaborado pelo autor

Como pode ser percebido, as análises variam entre variáveis quantitativas e qualitativas. Anteriormente, foi realizada a AED para as variáveis quantitativas; a seguir, será apresentada a AED para as variáveis qualitativas.

De imediato, observa-se que os **dados estão desbalanceados**, com um número muito maior de pacientes com alta em comparação aos que faleceram. A base possui 3.194 observações, sendo 2.962 de pacientes com **alta (92.74%)** e 232 de **óbitos (7.26%)**. Na Figura 15, nota-se que há mais homens internados do que mulheres. Embora isso não implique que os homens sejam necessariamente um grupo de risco, essa informação pode contribuir para discriminar as variáveis dentro dos modelos.

Figura 15: Gráfico de barras: Sexo por desfecho

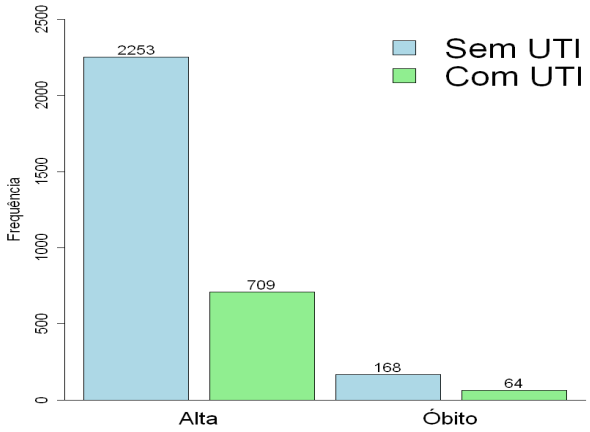


Fonte: Elaborado pelo autor

O gráfico da Figura 16 mostra que a grande maioria dos pacientes não foi internada na UTI. Quando analisado em conjunto com a proporção de altas e óbitos, esses dados,

isoladamente, não indicam fatores de risco evidentes, pois são superficiais e não trazem detalhes sobre as condições da internação ou características dos pacientes. No entanto, como no exemplo anterior, essas informações podem contribuir para a discriminação das variáveis dentro dos modelos.

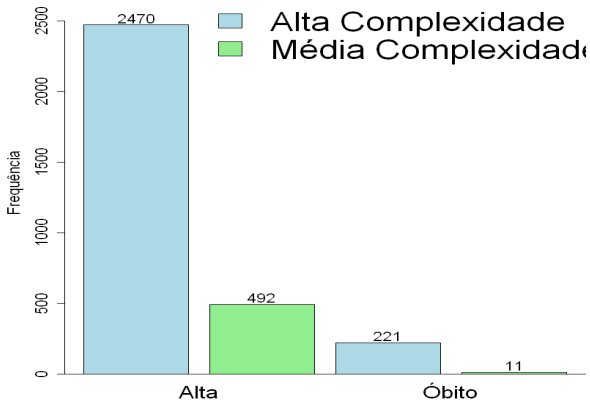
Figura 16: Gráfico de barras: Internações na UTI por desfecho



Fonte: Elaborado pelo autor

Ao analisar a proporção de complexidade dos procedimentos (Figura 17) realizados nos pacientes, observamos que a maioria pertence à classe de média complexidade. Isso indica que grande parte dos pacientes recebeu tratamentos de complexidade intermediária, o que pode impactar diretamente o tempo de internação e os custos associados. A distribuição dos pacientes entre as classes de complexidade também oferece *insights* sobre a gravidade das condições tratadas, contribuindo para uma melhor compreensão dos padrões de tratamento e das possíveis diferenças nos desfechos.

Figura 17: Gráfico de barras: Complexidade por desfecho

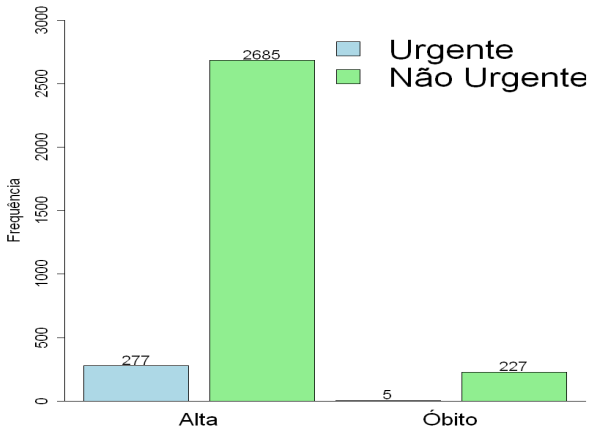


Fonte: Elaborado pelo autor

Por fim, não é surpresa que a grande maioria dos atendimentos por Infarto Agudo do Miocárdio (IAM) seja realizada em regime de urgência, como mostrado na Figura 18. Na

literatura, essa variável é frequentemente tratada como um fator de risco, e, a partir dos dados, ela realmente se apresenta como um bom discriminador para os modelos, ajudando a distinguir pacientes com maior risco de complicações e desfechos mais graves. Isso reforça a importância de considerar a urgência no atendimento como uma variável relevante para o desenvolvimento de modelos preditivos.

Figura 18: Gráfico de barras: Urgência por desfecho



Fonte: Elaborado pelo autor

4.2 Balanceamento dos dados

Seguindo a metodologia descrita no Capítulo 2, ao identificar dados desbalanceados, é necessário aplicar técnicas de balanceamento para adequar a base ao treinamento de modelos de classificação. **Optou-se pelo *oversampling***, pois a base contém cerca de 3.200 observações, com apenas 232 pertencendo à classe minoritária. O *undersampling* foi descartado para evitar a perda de informações relevantes, o que poderia comprometer o desempenho do modelo na classificação de novos dados.

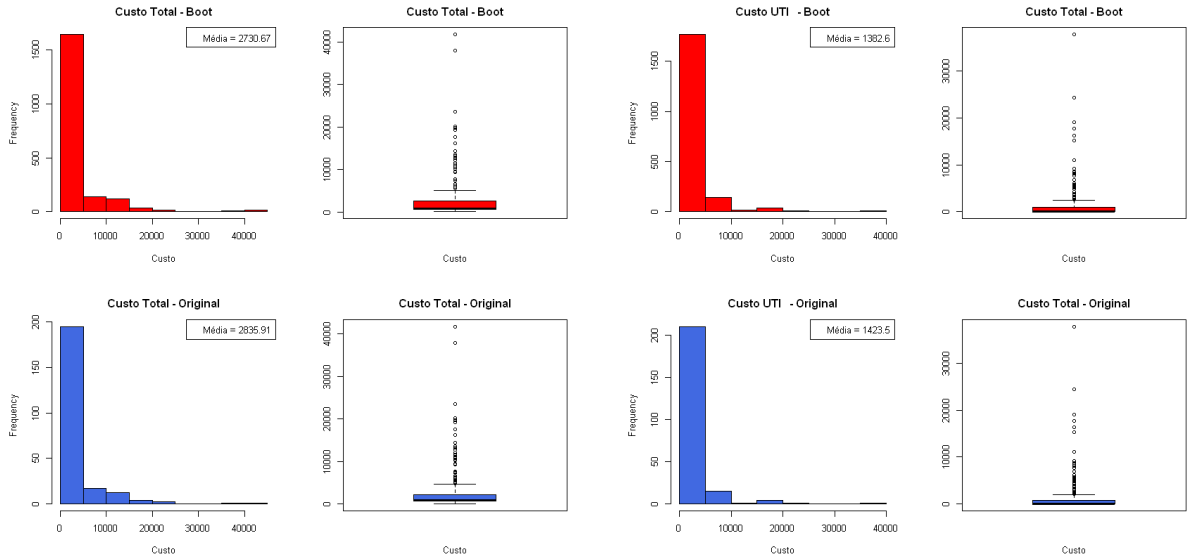
Para aumentar as observações da classe minoritária, escolheu-se a reamostragem com reposição (*bootstrap*), considerando o grande número de *outliers* na base, o que torna o algoritmo SMOTE menos adequado. A base foi balanceada na proporção de **60%** para pacientes com alta (**2.962 observações**) e **40%** para pacientes que faleceram, o que exigiu aumentar a classe minoritária para **1.974 observações**.

Após o balanceamento, é fundamental realizar uma nova Análise Exploratória de Dados para garantir que a amostra equilibrada continue representativa e adequada ao modelo.

4.2.1 Análise Exploratória da amostra *bootstrap*

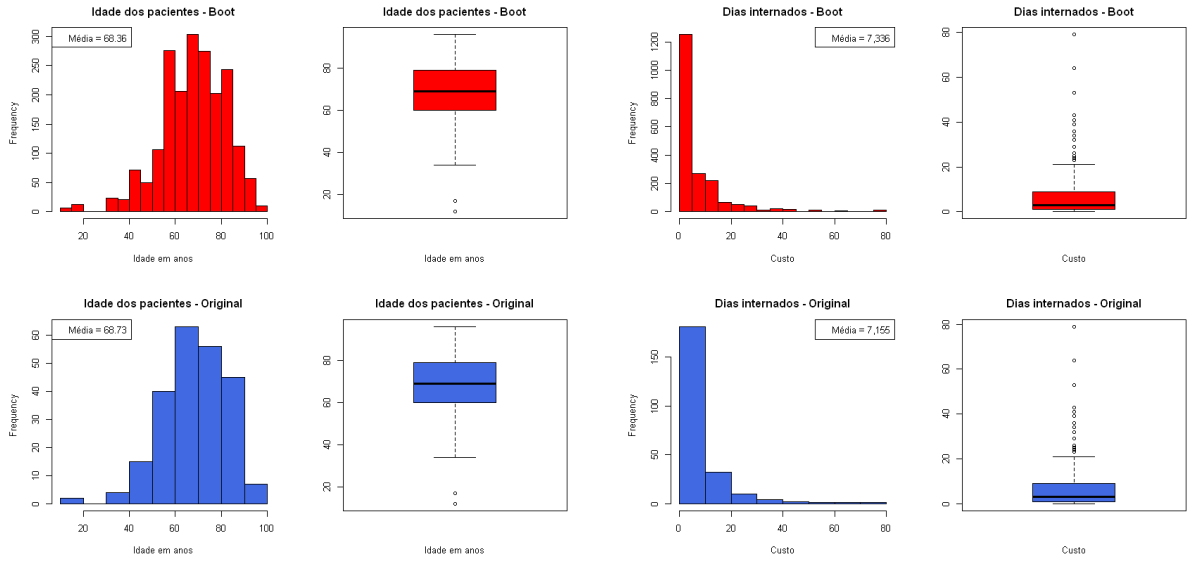
A análise realizada é breve e objetiva, consistindo em um estudo comparativo do comportamento das variáveis na amostra *bootstrap* em relação aos dados originais. Os resultados para as variáveis quantitativas são apresentados nas Figuras 19 e 20, enquanto as variáveis qualitativas estão sintetizadas na Tabela 5.

Figura 19: Distribuição Custo Total e UTI: *bootstrap* x original



Fonte: Elaborado pelo autor

Figura 20: Distribuição Idade e Permanencia: *bootstrap* x original



Fonte: Elaborado pelo autor

A reamostragem para as variáveis quantitativas mostrou-se representativa, reproduzindo os mesmos comportamentos observados nos dados originais e apresentando estatísticas semelhantes.

Tabela 5: Variáveis qualitativas: *bootstrap* x original

	Variável	***Bootstrap***	**Original**	***Prop. Boot***	**Prop. Orig**
Sexo	**Masculino**	1067	123	53.01%	54.05%
	Feminino	907	109	46.99%	45.95%
UTI	**Sim**	567	64	28.72%	27.59%
	Não	1407	168	71.28%	72.41%
Urgência	**Sem Urg.**	43	5	2.18%	2.15%
	Urgente	1931	227	97.82%	97.85%
Complexidade	**Média**	1872	102	95.26%	94.83%
	Alta	221	11	4.74%	5.17%

Fonte: Elaborado pelo autor.

Para as variáveis qualitativas, a reamostragem demonstrou ser representativa, aproximando - se das proporções originais de cada classe.

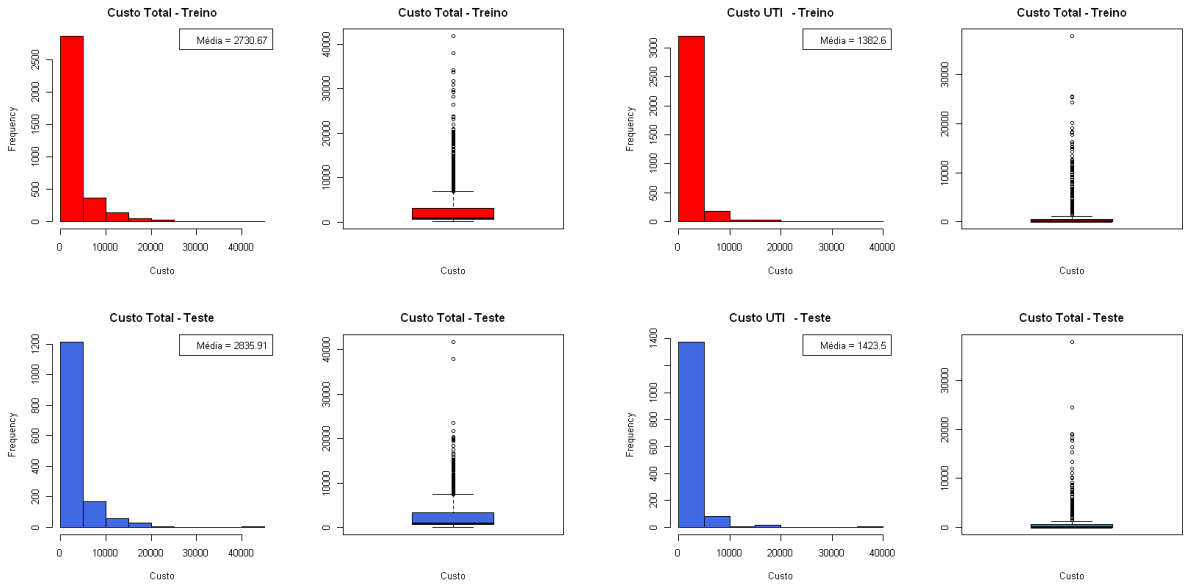
Assim, conclui-se que a reamostragem foi adequada, proporcionando uma base mais balanceada, o que permite prosseguir com a construção dos modelos de maneira mais eficaz.

5 Construção dos Modelos

Após o balanceamento dos dados, foram desenvolvidos os modelos de regressão logística, árvore de classificação e floresta aleatória no programa *RStudio*, utilizando 70% dos dados para treinamento e 30% para teste, juntamente com uma validação cruzada com 20 *folds*, respeitando a proporção das classes. Para validar a correta divisão, foi realizado um estudo comparativo, semelhante ao da subseção 4.2.1.

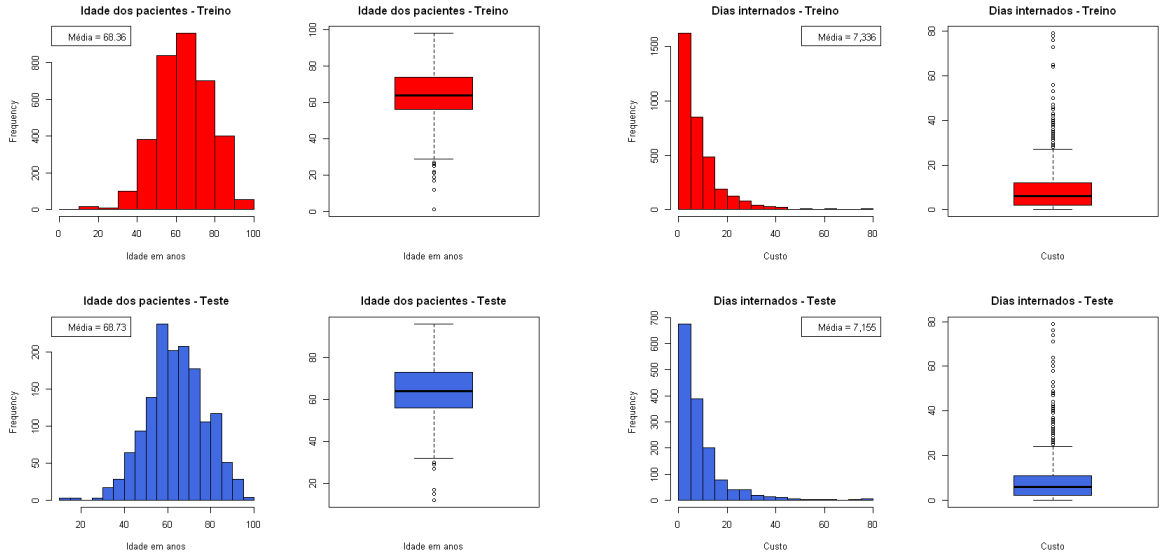
A proporção das classes nas bases de treino e teste manteve-se próxima à proporção original (60% alta e 40% óbito).

Figura 21: Distribuição Custo Total e UTI: *treinamento* x teste



Fonte: Elaborado pelo autor

Figura 22: Distribuição Idade e Permanencia: *treinamento* x teste



Fonte: Elaborado pelo autor

Para as variáveis quantitativas, a divisão demonstrou ser representativa, pois reproduziu os mesmos comportamentos observados nos dados originais e apresentou estatísticas semelhantes.

Tabela 6: Variáveis qualitativas: treinamento x teste

	Variável	**Treino**	**Teste**	**p. Treino**	**p. Teste**	**p. Orig**
Sexo	**Masculino**	1974	850	57.13%	57.40%	57.21%
	Feminino	1481	631	42.87%	42.60%	42.79%
UTI	**Sim**	877	399	25.38%	26.94%	25.85%
	Não	2578	1082	74.62%	73.06%	74.15%
Urgência	**Sem Urg.**	213	107	6.16%	7.22%	6.48%
	Urgente	3242	1374	93.84%	92.78%	93.52%
Complexidade	**Média**	3054	1288	88.40%	86.97%	87.97%
	Alta	401	193	11.60%	13.03%	12.03%

Fonte: Elaborado pelo autor.

Para as variáveis qualitativas, a divisão também se mostrou representativa, pois as proporções ficaram próximas das originais.

Dessa forma, a base de treino apresenta um bom desempenho para a construção dos modelos, enquanto a base de teste contém informações suficientes para validar os modelos, em relação aos valores populacionais.

A divisão de 70% para treino e 30% para teste foi escolhida para garantir que o modelo fosse treinado com um número significativo de dados, o que contribui para uma melhor generalização. Esse equilíbrio entre treino e teste possibilita uma validação robusta, com dados suficientes para avaliar a performance do modelo de forma confiável. Além disso, a aplicação de validação cruzada durante o treinamento, utilizando 20 *folds*, ajuda a reduzir o risco de superajuste (overfitting) e melhora o desempenho do modelo ao avaliar a sua performance em diferentes subconjuntos dos dados.

5.1 Regressão Logística

Para a construção do modelo de regressão logística, utilizou se as funções *glm* e o pacote *caret* para treino e validação cruzada. Todas as **variáveis categóricas** foram convertidas em ***factors*** de acordo com o padrão estabelecido na Tabela 4. Assim, a **variável resposta desfecho** tem como valor **1** o desfecho em **óbito** e **0** o desfecho em **alta**.

Com o objetivo de ajustar o melhor modelo, inicialmente foi construído um modelo com todas as covariáveis. Em seguida, utilizou-se o método *stepwise backward-forward* para selecionar as variáveis mais relevantes a serem incluídas no modelo final. Com essas variáveis, ajustou-se um modelo de regressão logística, cuja validação cruzada identificou

aquele com a melhor curva ROC. Abaixo, são apresentados os coeficientes do modelo, juntamente com seus erros padrão e $\mathbf{p}_{valor}$ do teste *Wald*:

Tabela 7: Coeficientes do Modelo de Regressão Logística para desfecho por IAM

Coeficientes	**Estimativa**	**Erro Padrão**	**pvalor**
Intercepto	3.4965	0.2999	2e-21
Custo Total	-4e-04	0.00003	2e-23
Dias UTI	0.059	0.0194	0.0023
Idade	-0.0392	0.003	3e-39
Permanencia	0.1195	0.0073	1e-60
Urgencia	-1.1389	0.2327	1e-07
Complexidade	3.9799	0.3208	2e-35

Fonte: Elaborado pelo autor.

5.1.1 Interpretação das *Odd Ratio*

Utilizando os valores da tabela 8 e um $\alpha = 0.05$ foram calculados os valores das OR e de seus respectivos intervalos de confiança. Segue abaixo o resultado.

Tabela 8: Intervalo de Confiança de 95% das OR

Coeficiente	**Lim. Inferior**	***Odd Ratio***	**Lim. Superior**
Intercepto	32.4129	33.0007	33.5884
Custo Total	0.9996	0.9996	0.9997
Dias UTI	1.0228	1.0608	1.0988
Idade	0.9557	0.9616	0.9675
Permanencia	1.1127	1.1269	1.1412
Urgência	-0.136	0.3202	0.7763
Complexidade	52.8825	53.5111	54.1398

Fonte: Elaborado pelo autor.

Observando a tabela 8, é possível extrair informações importantes sobre as covariáveis. Por exemplo, a partir da *Odd Ratio*, pode-se afirmar que um paciente internado com urgência apresenta uma redução de quase 68% nas chances de óbito em comparação com um paciente que não foi atendido com urgência. Além disso, observa-se que, a cada dia de internação na UTI, o paciente aumenta em quase 6% as chances de falecer, evidenciando a relação entre o tempo de permanência na unidade de terapia intensiva e o risco de desfechos fatais.

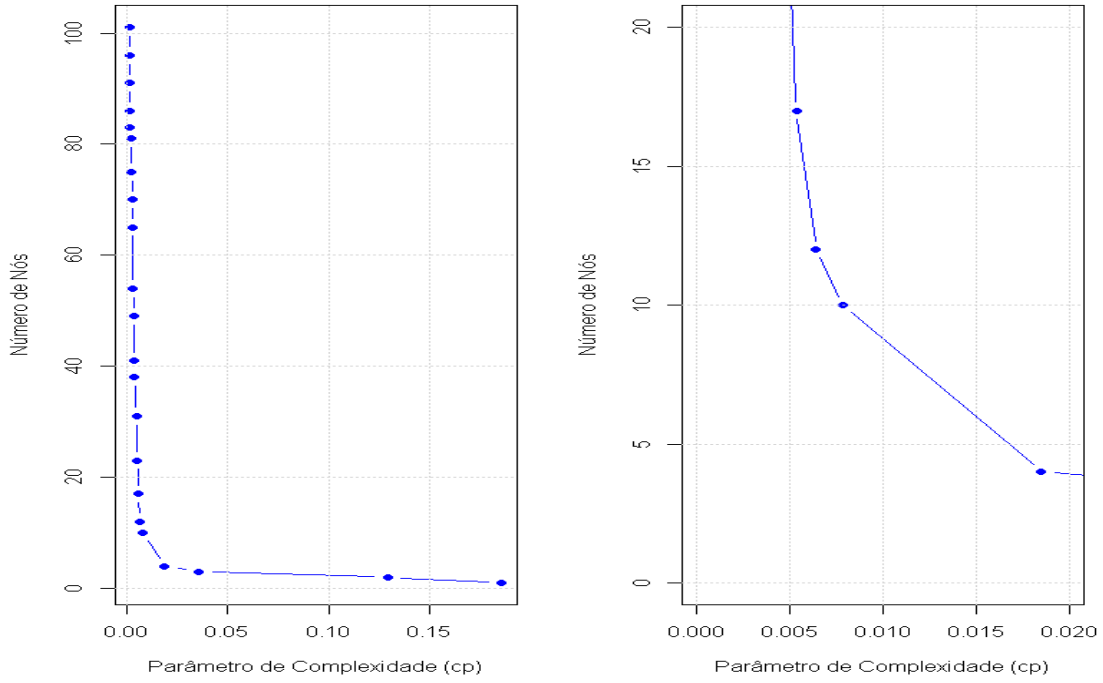
5.2 Árvore de Classificação

Para a construção da árvore de classificação, utilizou-se as funções do pacote *rpart* e, para o gráfico do modelo, o pacote *rpart.plot*. A validação cruzada e o treinamento do modelo foram realizados com as funções do pacote *caret*.

Inicialmente, foi construído um modelo com diferentes valores de *Complexity Price* (CP) para identificar o intervalo que equilibrasse complexidade e precisão, com o objetivo de gerar uma árvore interpretável. Com base nesse intervalo, o modelo final foi ajustado utilizando otimização por *tuning* de CP, além de uma limitação na profundidade da árvore, para evitar sobreajuste e garantir maior interpretabilidade.

Seguindo essas etapas, uma árvore de decisão inicial foi construída utilizando 50 valores de CP, variando de 0,001 a 0,1, com o índice Gini como critério de divisão. O modelo foi treinado para ajustar o CP, controlando o crescimento da árvore. O valor de CP que melhor equilibrou o número de nós e a precisão foi selecionado para garantir a interpretabilidade do modelo. O gráfico a seguir ilustra a relação entre o número de nós e o CP.

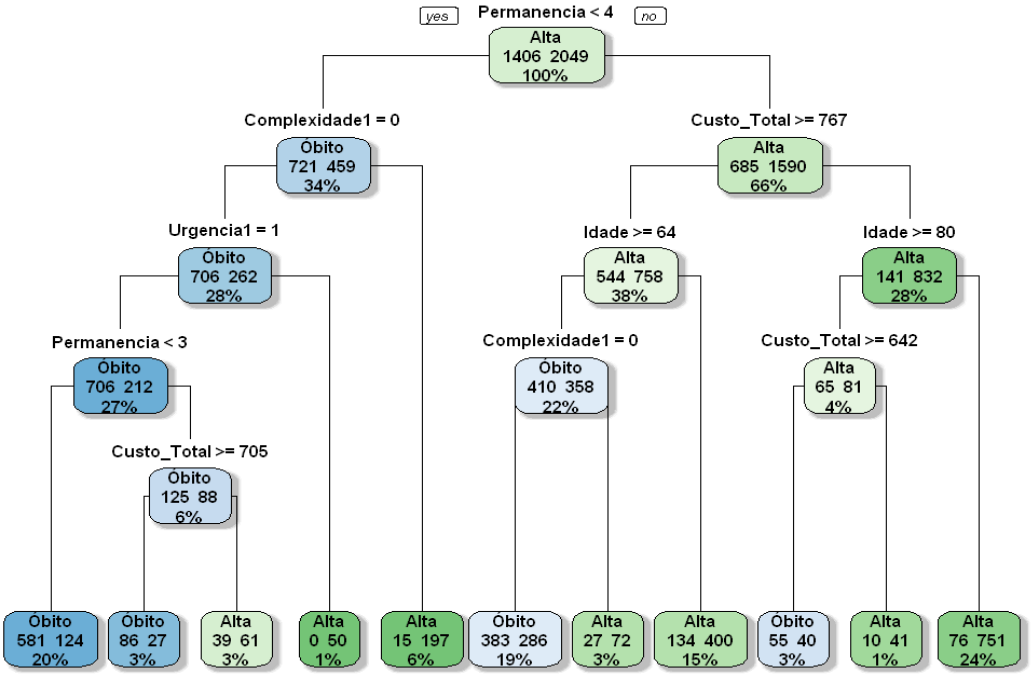
Figura 23: Variação do Número de Nós em Função de cp



Fonte: Elaborado pelo autor

Com o auxílio do gráfico, optou-se por construir um novo modelo utilizando 20 valores de CP, variando de 0.005 a 0.01, com o objetivo de obter uma árvore com até 5 níveis de profundidade. Abaixo está a árvore de decisão construída no *RStudio*:

Figura 24: Árvore de Classificação para o desfecho de internações por IAM



Fonte: Elaborado pelo autor

Guia para entender o modelo

Cada nó da árvore de decisão contém uma condição que divide as observações. O **fluxo à esquerda** indica que a observação **atende à condição**, enquanto o **fluxo à direita** indica o **contrário**.

Em cada nó, são apresentados três números: o **número central** representa a **frequência relativa** de observações, o **número à esquerda** indica a **frequência absoluta** de observações com **desfecho em óbito,** enquanto o **número à direita** representa a **frequência absoluta** das observações com **desfecho em alta**. As cores azul e verde representam visualmente a predominância dos desfechos das observações dentro dos nós, sendo **azul** referente ao **óbito** e **verde** referente à **alta**.

As **variáveis categóricas** foram transformadas em *factors*, conforme a tabela 4. Dessa forma, quando essas **variáveis** forem utilizadas na **divisão de um nó**, elas serão igualadas a 0 ou 1. Caso a condição seja a variável igual a 1, se a **observação** tiver valor 1, ela seguirá para o **fluxo da direita**; caso tenha valor 0, seguirá para o **fluxo da esquerda**.

5.2.1 Interpretação da árvore

Interpretar uma árvore de decisão pode parecer desafiador à primeira vista, mas ela continua relativamente simples de entender. A análise do desfecho pode ser feita a partir dos ramos, sendo mais eficiente focar nos principais ramos, cuja interpretação é direta.

De acordo com o critério de divisão por Gini, os primeiros nós representam as variáveis

que melhor separam as classes da variável de interesse ("desfecho"). Essas variáveis são as mais significativas no modelo e, dependendo da qualidade da separação, podem ser consideradas fatores de risco, o que é o objetivo principal deste estudo.

Interpretação do nó raíz

O nó raiz da árvore de decisão divide os pacientes com base no tempo de internação, separando-os em dois grupos: pacientes com menos de 4 dias de internação e pacientes com 4 dias ou mais. A divisão sugere certa discriminação entre os desfechos. Na partição de pacientes com **internação menor que 4 dias** (Permanência < 4 dias), observa-se uma **predominância de óbitos**, enquanto que na partição de pacientes com **4 dias ou mais de internação** (Permanência ≥ 4 dias), a maioria das observações corresponde a **pacientes com alta**. Dessa forma, pode-se sugerir que o tempo de internação tem alguma relação com os desfechos, indicando que internações mais curtas podem estar associadas ao óbito, mas sem que se possa afirmar uma causalidade direta entre internações curtas e a morte dos pacientes.

Interpretação de primeiros nós

Nos nós subsequentes, à direita, observa-se a divisão com base nos níveis de complexidade, enquanto à esquerda, temos a partição dos custos totais de internação superiores a 767 reais. O nó relacionado ao custo total demonstra uma boa discriminação entre as classes, pois a partição com custos inferiores a 767 reais apresenta uma predominância de desfechos de alta, enquanto a partição com custos superiores mostra uma distribuição mais equilibrada entre as classes. Dessa forma, pode-se concluir que este nó complementa de forma eficaz a divisão do nó raiz, ajudando a refinar a separação entre os desfechos.

Interpretação de um ramo

O nó à direita, relacionado à complexidade, se divide em outro nó à direita e em uma folha à esquerda, sendo este último considerado um ramo. O nó de nível de complexidade demonstrou uma boa capacidade de separação entre os desfechos, sugerindo que pacientes submetidos a procedimentos de complexidade média (Complexidade $= 0$) têm maior probabilidade de evoluir para óbito, em comparação com os pacientes que passaram por procedimentos de alta complexidade. Isso indica que pacientes com tempo de permanência inferior a 4 dias, que realizam procedimentos de alta complexidade, apresentam maiores chances de receber alta.

5.3 Floresta Aleatória

Para a construção da floresta aleatória, utilizou-se as funções do pacote *rpart*. A validação cruzada e o treinamento do modelo foram realizados com as funções do pacote *caret*.

Dentre os três modelos de Análise de Componentes (FA) avaliados, o modelo de FA destacou-se como o mais robusto em termos computacionais. Embora a otimização direta de hiperparâmetros seja vantajosa em diversas situações, sua aplicação pode se tornar inviável dependendo do volume de dados. Assim, optou-se por adotar os parâmetros previamente definidos no estudo da árvore desenvolvida anteriormente.

No estudo prévio, identificou-se que o valor do parâmetro de complexidade (CP) que melhor equilibra desempenho e interpretabilidade é 0,005. Durante a análise de árvores geradas com CP ligeiramente superior, observou-se que muitos nós próximos às folhas segmentavam grupos excessivamente pequenos, gerando nós com menos de 150 observações. Para otimizar o modelo, foi estabelecido um limite mínimo de observações por nó. Dessa forma, foram geradas 300 árvores para a construção do modelo final.

No modelo de FA, foi adotado o valor de CP igual a 0,005 e um critério mínimo de 125 observações por nó. Caso um nó possua menos de 125 observações, este será considerado uma folha. Adicionalmente, foi especificado no código que, em cada nó, fossem selecionadas aleatoriamente sete variáveis para análise.

5.3.1 Interpretação da floresta

Como mencionado anteriormente, o modelo não proporciona uma interpretação tão direta e objetiva quanto os outros dois, limitando-se a identificar as variáveis mais significativas para discriminar os dados e melhorar a acurácia. A importância de cada variável é apresentada na tabela abaixo.

Tabela 9: Importância das variáveis do modelo FA para o desfecho de internações por IAM

	Óbito	**Alta**	**Redução da precisão**	**Red. da impureza**
Sexo	0.0021	4e-05	8e-04	4.7308
Dias UTI	0.0143	0.0039	0.0082	15.6023
UTI	0.0014	6e-04	9e-04	1.9241
Custo Total	0.1816	0.0242	0.0883	229.1533
Custo UTI	0.0604	7e-04	0.0249	32.5057
Idade	0.1031	0.0189	0.0531	139.8505
Permanencia	0.2835	0.1334	0.1943	223.2556
Urgencia	0.026	0.009	0.0159	45.1482
Complexidade	0.0357	0.0226	0.028	92.963

Fonte: Elaborado pelo autor

A interpretação do modelo de Floresta Aleatória (RF) construído revela que as variáveis **Permanência, Custo Total, Idade e Dias na UTI** apresentam maior relevância na previsão do desfecho de pacientes internados por Infarto Agudo do Miocárdio (IAM), uma vez que apresentam elevados valores tanto na **redução da precisão** quanto na **redução da impureza**. Esses resultados indicam que tais variáveis desempenham um papel crucial na acurácia do modelo e na sua capacidade de discriminar adequadamente os diferentes desfechos. Em contrapartida, as **variáveis Sexo, UTI** e **Urgência** apresentam valores mais baixos em ambas as métricas, sugerindo uma **contribuição menos significativa** para o desempenho do modelo. Dessa forma, o modelo destaca que fatores como a duração da internação (Permanência), os custos totais (Custo Total), a idade do paciente e o tempo de permanência na UTI são os principais determinantes na previsão do desfecho, seja ele alta ou óbito.

5.4 Métricas dos modelos

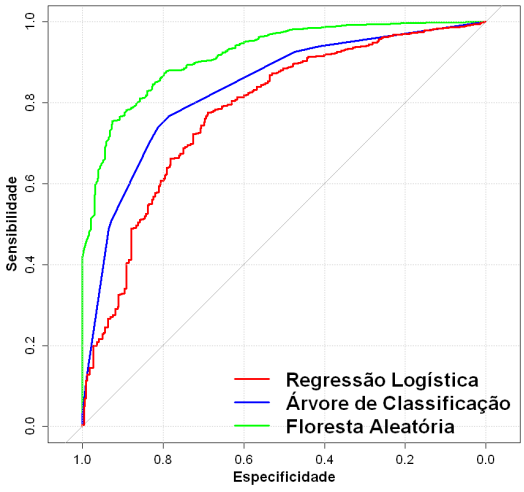
A seguir é apresentado as principais métricas de classificação dos três modelos e assim comparar o desempenho de cada um deles.

Tabela 10: Métricas de classificação dos modelos construídos

	Métricas					
Modelo	**Acurácia**	**Sensibil.**	**Especific.**	**Precisão**	**Kappa**	**AUC**
RL	0.7454	0.610	**0.830**	0.691	0.450	0.7967
AC	0.765	**0.750**	0.774	0.6741	0.513	0.8287
FA	**0.815**	0.706	0.883	0.789	**0.601**	**0.9081**

Fonte: Elaborado pelo autor

Figura 25: Curva ROC dos três modelos



Fonte: Elaborado pelo autor

Um aspecto relevante na comparação dos modelos é o tempo de execução de cada código. Os testes foram realizados em uma máquina equipada com um processador AMD *Ryzen* 5 3550H e 24 GB de memória RAM.

Tabela 11: Tempo de execução dos modelos em segundos

Modelo	**Tempo de Processamento**	**Tempo Real**
Regressão Logística	0.53	0.78
Árvore de Classificação	2.12	3.64
Floresta Aleatória	113.07	161.89

Fonte: Elaborado pelo autor

Os resultados apresentados na Tabela 10 mostram que, embora a base tenha sido ajustada para reduzir o desbalanceamento original, a desproporção entre as classes afeta consideravelmente o poder preditivo do modelo. Essa desproporção impacta a confiabilidade de métricas como a especificidade e, em menor grau, a acurácia.

Nos dados de internações hospitalares, priorizar a sensibilidade é essencial, pois falsos negativos (FN) – casos em que pacientes de alto risco são classificados como de baixo risco – podem levar à omissão de tratamentos ou monitoramento intensivo, aumentando o risco de óbito. Já falsos positivos (FP), apesar de gerarem custos adicionais, são mais aceitáveis, pois garantem a segurança do paciente.

Dado a leve desproporção das classes (60% alta e 40% óbito), o coeficiente *Kappa* é uma métrica relevante, visto que ela considera a desigualdade das classes. A curva ROC também complementa a avaliação, permitindo comparar a capacidade discriminativa dos modelos. Adiante é feito uma discussão individual dos modelos e por fim a comparação entre eles.

A Regressão Logística (RL) apresentou o pior desempenho geral. Sua sensibilidade (0.610) foi a menor entre os modelos, indicando dificuldade em identificar corretamente os casos positivos, o que resulta em um número maior de falsos negativos, algo crítico neste contexto. O coeficiente Kappa de 0.450 reforça essa limitação, sugerindo baixa concordância ajustada para o desbalanceamento dos dados. Embora a especificidade (0.830) seja alta, isso não compensa a baixa sensibilidade, já que minimizar falsos negativos é mais importante do que evitar falsos positivos.

A Árvore de Classificação (AC) teve desempenho intermediário, destacando-se pela maior sensibilidade (0.750), o que a torna a melhor escolha para reduzir falsos negativos. Esse modelo, portanto, é mais alinhado a cenários em que a prioridade é identificar todos os casos positivos possíveis, mesmo ao custo de um aumento nos falsos positivos. O coeficiente Kappa (0.513) sugere uma melhora em relação à RL, mas ainda inferior à Floresta Aleatória. A especificidade (0.774) foi menor, indicando que esse modelo pode ser menos eficiente na classificação de casos negativos.

A Floresta Aleatória (FA) demonstrou o melhor desempenho geral, com acurácia de 0.815 e o maior coeficiente Kappa (0.601), o que reflete boa performance em cenários com desbalanceamento. Apesar de sua sensibilidade (0.706) ser inferior à da AC, ela é suficiente para um bom desempenho na identificação de casos positivos, enquanto sua alta AUC (0.9081) evidencia excelente capacidade discriminativa entre as classes. Isso faz da FA o modelo mais equilibrado para o conjunto de dados, conseguindo lidar melhor com a desproporção das classes.

Dessa forma, enquanto a Floresta Aleatória é a escolha mais equilibrada, sendo robusta em diferentes métricas, a Árvore de Classificação se destaca como uma alternativa válida em cenários onde minimizar falsos negativos é prioritário. A Regressão Logística, embora simples e interpretável, foi a menos indicada para este caso específico devido ao seu baixo desempenho em métricas críticas, especialmente a sensibilidade.

6 Considerações Finais

De forma geral, os três modelos analisados apresentaram desempenhos influenciados pelo desbalanceamento dos dados, com a **Floresta Aleatória** se destacando pelo **melhor desempenho** preditivo, evidenciado pelos maiores valores de AUC e Kappa. No entanto, sua baixa interpretabilidade pode limitar sua aplicação em contextos nos quais explicações claras são essenciais para a tomada de decisão. A Árvore de Classificação, por sua vez, demonstrou um equilíbrio satisfatório entre desempenho e interpretabilidade, sendo mais adequada para situações que exigem transparência nas decisões. Já a **Regressão Logística** apresentou o **menor desempenh**o, especialmente em termos de sensibilidade, mas se mantém relevante como modelo de base devido à sua simplicidade e facilidade de interpretação.

Ao analisar os fatores de risco identificados pelos modelos, percebe-se que, embora a Regressão Logística forneça *Odds Ratios* que indicam a direção das associações, a precisão desses valores pode ser limitada pela incapacidade do modelo em capturar interações não lineares presentes nos dados. As variáveis **menos significativas** nos três modelos foram as mesmas: **sexo e internação na UTI**, indicando que esses fatores não apresentam impacto substancial no desfecho de óbito dentro deste conjunto de dados. Por outro lado, as variáveis mais significativas identificadas foram o tempo de permanência, o custo total da internação e a idade do paciente, cujas interpretações divergiram entre os modelos, refletindo as diferentes abordagens adotadas por cada um.

Na Regressão Logística, o tempo de permanência sugere que a cada dia adicional de internação, a chance de óbito aumenta em aproximadamente 12%. Em contrapartida, a Árvore de Classificação aponta que internações mais curtas estão associadas a maior risco de óbito. A Floresta Aleatória, por sua vez, identificou o tempo de permanência como uma das variáveis mais relevantes, contribuindo tanto para a precisão quanto para a redução da impureza nos nós da árvore. Em relação ao custo total da internação, na Regressão Logística, a **cada unidade adicional** de custo, a chance **de óbito diminui** em 0,0004%. Na Árvore de Classificação, custos superiores a 640 estão associados a um maior risco de óbito, enquanto na Floresta Aleatória, o custo total se destacou como a segunda variável mais importante. Por fim, a interpretação da variável **idade** foi inversa entre os modelos: na Regressão Logística, um aumento na idade **reduz a chance de óbito em cerca de 40%**, enquanto na Árvore de Classificação, pacientes mais idosos apresentam maior risco de óbito.

Essas divergências nas interpretações indicam que as covariáveis possuem interações não lineares que afetam os desfechos analisados. Embora a Regressão Logística seja o modelo mais interpretável, sua incapacidade de capturar essas interações não lineares limita sua adequação para uma análise mais precisa. Dessa forma, considerando tanto o uso prático quanto a necessidade de interpretabilidade para auxiliar na tomada de decisões,

a Árvore de Classificação se mostrou a melhor escolha, apesar de sua precisão ser inferior à da Floresta Aleatória.

REFERÊNCIAS

BRAGA, A. C. **Curva ROC:** Aspectos funcionais e aplicações. 2000. Tese (Doutorado em Estatística) - Universidade do Minho, Braga, 2000. Disponível em: $https://repositorium.sdum.uminho.pt/bitstream/1822/195/1/tese_doutACB.pdf$. Acesso em: 13 set. 2024.

BRASIL. Ministério da Saúde. **Infarto**. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/i/infarto. Acesso em: 28 out. 2024.

BREIMAN, L. et al. **Classification and Regression Trees**. Filadélfia, PA, USA: Chapman & Hall/CRC, 1984.

CRAMER, J. S. **The Origins of the Logistic Regression**. Tinbergen Institute Working Paper, Amsterdã, Nº 2002-119/4, 2002. Disponível em: https://papers.tinbergen.nl/02119.pdf. Acesso em 04 jun. 2024.

FACELI, K. et al. **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC. 2011. Acesso em: 21 nov. 2024.

HASTIE, T.; Tibshirani, R.; Friedman, J.**The elements of statistical learning:** Data mining, inference, and prediction, second edition. 2. ed. Nova Iorque, NY, USA: Springer, 2009.

HOSMER, D.; LEMESHOW, S.; STURDIVANT, X. **Applied Logistic Regression**. 3th ed. Canada: John Wiley Sons, 2013.

IZBICKI, R.; DOS SANTOS, T. M. **Aprendizado de máquina:** uma abordagem estatística. [s.l.] Rafael Izbicki, 2020. Disponível em: http://www.rizbicki.ufscar.br/AME.pdf. Acesso em: 24 mai. 2024.

LOGUNOVA, I. Guide to random forest classification and regression algorithms. Disponível em: https://serokell.io/blog/random-forest-classification. Acesso em: 22 nov. 2024.

MORETTIN, P. A.; SINGER, J. M. **Estatística e ciência de dados**. São Paulo: [s.l.], abr. 2021. Versão preliminar.

NAGAURI, M. R. Guide to ensemble methods: Bagging vs Boosting. Disponível em: https://analyticsindiamag.com/ai-mysteries/guide-to-ensemble-methods-bagging-vs-boosting/. Acesso em: 22 nov. 2024.

OLIVEIRA, G. et al. Estatística Cardiovascular – Brasil 2023. **Arquivos Brasileiros de Cardiologia**, v. 121, n. 2, e20240079, mar. 2024. Acesso em 28 out. 2024.

PAIXÃO, G. et al. Machine Learning na Medicina: Revisão e Aplicabilidade. **Arquivos brasileiros de cardiologia**, v. 118, n. 1, p. 95–102, 2022. Disponível em: https://www.scielo.br/j/abc/a/WMgVngCLbYfJrkmC65VFCkp/?format=pdf&lang=pt. Acesso em: 8 out. 2023.

SCIKIT LEARN. **Cross-validation:** evaluating estimator performance. Disponível em: $https://scikit-learn.org/stable/modules/cross_validation.html$. Acesso em: 22 nov. 2024.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineiração de dados:** com aplicações em R. 1. ed. Rio de Janeiro: Elsevier, 2016. Acesso em: 22 jul. 2024.