Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Bruna Freitas dos Santos

Análise comparativa de modelos preditivos e de classificação para avaliação da função pulmonar na Fibrose Cística

Maringá,
28 de agosto de 2025

Bruna Freitas dos Santos

Análise comparativa de modelos preditivos e de classificação para avaliação da função pulmonar na Fibrose Cística

Dissertação apresentada à Universidade Estadual de Maringá, como parte das exigências do Programa de Pós-Graduação em Bioestatística, área de concentração em Bioestatística, para a obtenção do título de Mestre.

Universidade Estadual de Maringá - UEM

Orientador: Willian Luís de Oliveira

Maringá,
28 de agosto de 2025

Resumo

O presente estudo teve como objetivo principal propor técnicas de *Aprendizado de máquina* para prever o volume expiratório forçado no primeiro segundo (VEF1), com isso foi analisado fatores que influenciam o VEF1 em pacientes com Fibrose Cística (FC), como capacidade vital forçada (CVF), número de consultas, idade, IMC, sexo, raça, região e genótipo. Ainda, além de considerar o VEF1 numericamente, o mesmo foi categorizado em 3 cenários diferentes, sendo o primeiro, segundo e terceiro quartil. As técnicas propostas seguiram as abordagens de modelos de predição e modelos de classificação. Ainda, foi feito uma análise descritiva nos dados e a mesma mostrou uma correlação forte entre a capacidade vital forçada (CVF) e o VEF1, com a CVF se destacando como principal preditor. Como resultado obtido, o modelo floresta aleatória obteve o melhor desempenho em ambos tipos de abordagens, enquanto que a árvore de regressão e o MLG tiveram desempenho inferior. O estudo indicou que o método floresta aleatória é uma ferramenta preditiva eficaz para prever VEF1 e que a CVF é um determinante considerável, além do modelo generalizado logístico permanecer como alternativa relevante em casos de modelos de classificação.

Palavras-chaves: Aprendizado de máquina, Classificação, Fibrose cística, predição, volume expiratório forçado.

Abstract

The main objective of this study was to propose machine learning techniques to predict forced expiratory volume in the first second (FEV1). Factors influencing FEV1 in patients with cystic fibrosis (CF) were analyzed, such as forced vital capacity (FVC), number of consultations, age, BMI, gender, race, region, and genotype. In addition to considering FEV1 numerically, it was categorized into three different scenarios: first, second, and third quartiles. The proposed techniques followed the approaches of prediction models and classification models. Furthermore, a descriptive analysis of the data was performed, which showed a strong correlation between forced vital capacity (FVC) and FEV1, with FVC standing out as the main predictor. As a result, the random forest model performed best in both types of approaches, while the regression tree and GLM performed poorly. The study indicated that the random forest method is an effective predictive tool for predicting FEV1 and that FVC is a considerable determinant, in addition to the generalized logistic model remaining a relevant alternative in cases of classification models.

Key-words: Classification, Cystic fibrosis, machine learning, prediction, forced expiratory volume.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

FC Fibrose Cística

RTFC "Regulador da condutância transmembranar da fibrose cística"

KNN "k-vizinhos mais próximos"

DRC Doença renal crônica

SM Síndrome metabólica

MA Maranhão

IRT Tripsinogênio imunorreativo

SRAG Síndrome Respiratória Aguda Grave

GBEFC Grupo Brasileiro de Estudos de Fibrose Cística

REBRAFC Registro Brasileiro de Fibrose Cística

AM Aprendizado de máquina

RS Rio Grande do Sul

VEF1 Volume expiratório forçado

CVF Capacidade vital forçada

IMC Índice de massa corporal

MLG Modelos lineares generalizados

SQR Soma de quadrados dos resíduos

MSE Erro quadrático médio

RMSE Raiz do erro quadrático médio

MAE Erro absoluto médio

VP Valores positivos

VN Valores negativos

FP Falso positivo

FN Falso negativo

VPP Valor preditivo positivo

VPN Valor preditivo negativo

k Coeficiente Kappa de Cohen

AUC Área sob a curva

ROC Característica de Operação do Receptor

OR Razões de chances

Sumário

Capítulo 1

Introdução

1.1 Contextualização do tema

1.1.1 Histórico

A primeira descrição das características importantes da fibrose cística (FC) surgiu em 1905, quando o patologista Landsteiner a denominou "fibrose cística do pâncreas", onde foi descrito que essa condição era uma doença que afeta o pâncreas exócrino, sem envolvimento das ilhotas de Langerhans (Campos et al., 1996). Landsteiner também foi o primeiro a relatar achados anatomopatológicos em recém-nascidos que faleceram no quinto dia de vida devido ao íleo meconial.

Em 1936, Guido Fanconi e colaboradores relataram um caso de uma criança com sintomas gastrointestinais e pulmonares que, apesar da semelhança, à síndrome celíaca, apresentava alterações pancreáticas diferentes da doença celíaca clássica. A caracterização científica da FC ocorreu com o estudo de Andersen (1938), da Universidade de Columbia. Andersen analisou clinicamente e em necropsias 49 casos de crianças e lactentes, oferecendo a primeira descrição detalhada dos sintomas e das alterações orgânicas da mesma. Ela também observou a associação entre lesões pulmonares e pancreáticas e identificou o íleo meconial como uma das primeiras manifestações da doença. Andersen ainda consolidou a FC como uma doença multissistêmica e de provável herança genética recessiva.

Com base na descrição de Andersen, Farber (1944), sugeriu o termo "mucoviscidose"para enfatizar a espessura e viscosidade das secreções que obstruíam ductos pancreáticos e vias aéreas. Di Sant'Agnese et al. (1953) identificaram altos níveis de sódio e cloro no suor dos pacientes, descoberta que levou a criação do teste do suor. O método de Gibson and Cooke (1959), permanece amplamente utilizado para esse teste, que consiste na estimulação da pele com iontoforese de pilocarpina.

Nos anos seguintes, o diagnóstico e o tratamento da FC foram aprimorados, embora os avanços bioquímicos fossem mais lentos que os clínicos. Em Quinton (1983), foi identificado

que o epitélio dos pacientes era impermeável aos íons cloreto, sugerindo uma falha em canais de transporte desse íon. Essa descoberta levou à compreensão de que o déficit de água nas secreções era responsável pela obstrução das glândulas exócrinas e pela disfunção de múltiplos órgãos (de Faria, 2007).

Em Wainwright et al. (1985), a localização do gene da FC foi identificada no cromossomo 7. Quatro anos depois, em 1989, pesquisadores descobriram o gene Regulador da condutância transmembranar da fibrose cística (RTFC), cuja mutação causa a FC, e confirmaram que a proteína resultante, um canal de cloro defeituoso, localiza-se na membrana apical do epitélio (Bear et al., 1992). Essa descoberta não só ampliou o entendimento sobre a doença, mas ainda abriu caminho para o desenvolvimento de terapias inovadoras, direcionadas à cura e não apenas ao controle dos sintomas.

1.1.2 Genética

O gene responsável pela FC codifica a proteína RTFC, que é composta por 1480 aminoácidos. A RTFC atua de forma principal como um canal de cloro ativado pelo monofosfato de adenosina cíclica (AMPc), mas ainda participa da regulação de outros canais iônicos, desempenhando um papel essencial na composição do fluido da superfície epitelial.

Desde a identificação do gene da FC em 1989, foram descobertas mais de 1800 mutações que afetam a produção da RTFC por diferentes mecanismos moleculares, causando um déficit funcional da proteína, seja em quantidade ou qualidade. Contudo haja um grande número de mutações, muitas são raras. A mutação mais comum e extensivamente estudada é a F508del (anteriormente chamada de $\Delta F508$), que se caracteriza pela ausência de três nucleotídeos consecutivos (uma citosina e duas timinas), resultando na perda de um resíduo de fenilalanina (F) na posição 508 da RTFC. A $F508del$ interfere diretamente no processamento da proteína, impedindo seu funcionamento adequado.

1.1.3 Aspectos epidemiológicos

Aproximadamente uma em cada 25 pessoas na população é portadora do gene defeituoso da FC, que se manifesta quando uma criança herda uma cópia mutada do gene RTFC de ambos os pais.

A prevalência da FC varia conforme a etnia, com uma taxa de 1 em 2.000 a 1 em 5.000 em caucasianos na Europa, Estados Unidos e Canadá, 1 em 15.000 entre negros americanos, e 1 em 40.000 na Finlândia, sendo rara em asiáticos e africanos. No Brasil, a prevalência estimada na região Sul é mais próxima dos padrões da população caucasiana da Europa Central, enquanto diminui nas regiões Sudeste e Norte. Porém, a ausência de estudos abrangentes de triagem neonatal e dados epidemiológicos completos dificulta a estimativa precisa da incidência no país.

Nos primeiros registros entre as décadas de 1930 e 1940, a FC era considerada uma doença infantil, com 80% das crianças afetadas falecendo no primeiro ano de vida e poucas sobrevivendo além dos cinco anos. Com os avanços no diagnóstico e nas intervenções terapêuticas nas últimas três décadas, a expectativa de vida vem aumentando, embora ainda 15% a 20% das crianças falecem antes dos dez anos e a sobrevivência após os 30 anos ainda é um desafio. Pacientes tratados em centros especializados em FC, com equipes treinadas para acompanhamento e prevenção de complicações, com suporte nutricional adequado, terapias antibióticas precoces e intensivas apresentam um prognóstico melhor, principalmente quando o diagnóstico e o tratamento ocorrem de forma precoce, antes de ocorrerem danos pulmonares.

1.1.4 Diagnóstico

Para diagnosticar a FC após uma triagem neonatal positiva, o protocolo adotado no Brasil recomenda inicialmente a quantificação dos níveis de tripsinogênio imunorreativo (IRT) em duas amostras, sendo a segunda realizada até o $30ž$ dia de vida. Caso ambas as dosagens apresentem níveis elevados, realiza-se o teste do suor para confirmação, considerado o padrão-ouro para o diagnóstico. No teste, concentrações de cloro ($60mEq/L$ ou mais, dependendo da técnica utilizada) em duas amostras confirmam o diagnóstico de fibrose cística.

Além do teste do suor, a confirmação pode ser pela identificação de duas mutações do gene RTFC associadas à doença ou pela realização de testes de função da proteína RTFC. Além disso, vale lembrar que a triagem neonatal não confirma ou exclui o diagnóstico de FC de forma definitiva, mas identifica recém-nascidos com maior risco para a doença, sendo possível a ocorrência de resultados falso-positivos.

Com isso, o diagnóstico da doença pode ser sustentado por achados clínicos típicos, como sintomas pulmonares e gastrintestinais característicos, além de histórico familiar da doença, corroborado por exames laboratoriais específicos.

1.1.5 Grupo Brasileiro de Estudos de Fibrose Cística (GBEFC)

O Grupo Brasileiro de Estudos de Fibrose Cística (GBEFC), fundado em 5 de novembro de 2003, é uma organização sem fins lucrativos formada por profissionais de saúde especializados na área. Entre suas principais atividades estão a promoção de pesquisas, o treinamento de profissionais, o apoio na criação de centros de tratamento, a organização de congressos sobre FC e a colaboração com o Ministério da Saúde para estabelecer um protocolo nacional de atendimento à doença. Desde 2009, o grupo mantém um banco de dados de pacientes com fibrose cística chamado Registro Brasileiro de Fibrose Cística (REBRAFC), administrado pelo Laboratório de Sistemas Integráveis da Escola Politécnica da Universidade de São Paulo.

Com essas informações, o GBEFC também elabora relatórios anuais, que ficam disponíveis no site do GBEFC (www.gbefc.org.br), na qual contêm estatísticas descritivas sobre

dados demográficos, diagnósticos e tratamentos de pacientes com a doença, além de serem atualizados anualmente.

1.2 Revisão da literatura

O estudo de Reis and Damaceno (1998) teve como objetivo realizar uma revisão abrangente sobre a fibrose cística (FC), motivada pelos avanços no tratamento e pela necessidade de maior conhecimento da doença entre pediatras. A revisão baseou-se em trabalhos científicos da literatura médica internacional e abordou de forma detalhada aspectos como introdução, histórico, genética, fisiopatologia, microbiologia das infecções pulmonares, manifestações clínicas, critérios diagnósticos, diagnóstico diferencial, tratamento e prognóstico.

O estudo de Ribeiro et al. (2002), conduzido por Jose Dirceu Ribeiro, Maria Ângela G. de O. Ribeiro e Antonio Fernando Ribeiro, relatou que nos últimos 70 anos a fibrose cística (FC) tornou-se a principal doença genética potencialmente letal entre a população branca. Embora afete múltiplos órgãos e sistemas, sua expressão clínica varia entre os indivíduos, com manifestações predominantemente pulmonares e digestivas. O trabalho teve como objetivo atualizar o pediatra geral sobre os principais aspectos da FC, abordando incidência, fisiopatologia, manifestações clínicas, diagnóstico e tratamento, com base em uma revisão sistemática de 79 artigos. Apesar de ainda não haver cura, os avanços recentes na compreensão da etiologia e fisiopatologia da doença têm possibilitado novas abordagens terapêuticas e contribuído para o aumento da sobrevida e da qualidade de vida dos pacientes.

No trabalho de Lemos et al. (2004), conduzido pelo Núcleo de Pesquisa em Pneumologia (NUPEP) no Centro de Referência de Fibrose Cística da Bahia, avaliou aspectos clínicos e espirométricos de 28 pacientes adultos com fibrose cística (FC). A média de idade foi de $31,1$ anos, com $53,7\%$ de negros ou mulatos e índice de massa corpórea médio de $18,7\,\mathrm{kg/m^2}$. A presença de *Pseudomonas aeruginosa* foi identificada em 43% dos casos, e os valores médios de capacidade vital forçada (CVF) e volume expiratório forçado no primeiro segundo (VEF1) corresponderam a $58,9\%$ e $44,1\%$ do previsto, respectivamente. Pacientes colonizados por *P. aeruginosa* apresentaram maior comprometimento da função pulmonar. O estudo destaca a importância de considerar o diagnóstico de FC em adultos com infecções respiratórias recorrentes e outras complicações pulmonares.

O estudo de Furgeri et al. (2006) investigou haplótipos em núcleos familiares de pacientes com fibrose cística (FC) e avaliou o uso de polimorfismos no diagnóstico pré-natal e pré-implantação, correlacionando-os à mutação $\Delta F508$. A análise dos polimorfismos GATT, MP6-D9, TUB09 e TUB18, realizada por PCR e digestão enzimática, identificou nove haplótipos em 39 cromossomos. O haplótipo $6, +; G; C$ foi o mais frequente e apresentou forte associação com a mutação $\Delta F508$. Em 43% das famílias analisadas, foi identificado ao menos um polimorfismo informativo para diagnóstico pré-natal ou pré-implantação. Esses marcadores

mostraram-se úteis para o acompanhamento da transmissão de alelos mutados em famílias com FC.

Em 2008, o estudo de Rosa et al. (2008) revisou a literatura sobre os aspectos clínicos e nutricionais da fibrose cística (FC). A doença, causada por mutação no gene regulador da condutância transmembrana, leva à produção de secreções espessas que obstruem pulmões, pâncreas e ductos biliares. Muitos pacientes apresentam insuficiência pancreática, resultando em má absorção de nutrientes, especialmente proteínas e lipídios, além de complicações gastrointestinais como prolapso retal, obstrução intestinal, constipação e cirrose hepática. Diante da complexidade e variabilidade clínica da FC, o estudo destaca a importância de uma abordagem multidisciplinar para o manejo da doença. O tratamento inclui manutenção nutricional, fisioterapia respiratória, uso de mucolíticos e antibióticos, além de dietas ricas em lipídios e proteínas, com suplementação de minerais e vitaminas lipossolúveis.

No estudo de de Cássia Firmida et al. (2011) realizou uma revisão da literatura sobre a fisiopatologia da fibrose cística (FC) e sua relação com as manifestações clínicas. Os autores destacam que os avanços no entendimento da doença têm contribuído para melhorias significativas no diagnóstico, no acompanhamento e no tratamento dos pacientes.

1.2.1 Técnicas de Aprendizado de máquina em trabalhos da área da saúde

O estudo de Bittencourt et al. (2024) teve como objetivo desenvolver um modelo preditivo para identificar a síndrome metabólica (SM) em pacientes com doença renal crônica (DRC). Trata-se de um estudo transversal prospectivo realizado entre janeiro de 2018 e julho de 2020 em um centro de referência em São Luís (MA), envolvendo 196 adultos, com média de idade de $44,7$ anos, predominância feminina ($71,9\%$) e alta prevalência de sobrepeso ($69,4\%$). O algoritmo de aprendizado de máquina utilizado foi o k-vizinhos mais próximos (KNN), que apresentou acurácia e sensibilidade de 79%, especificidade de 80% e área sob a curva ROC de $0,79$, indicando bom desempenho. A SM associou-se a maior circunferência da cintura, pressão arterial elevada e glicemia de jejum aumentada. O estudo conclui que o KNN é uma ferramenta de triagem de baixo custo e viável para identificar o risco de SM em pacientes com DRC, favorecendo intervenções precoces e prevenção de complicações cardiovasculares.

A pesquisa de da Silva Bezerra and de Almeida (2024) teve como objetivo desenvolver modelos preditivos para estimar o óbito por Síndrome Respiratória Aguda Grave (SRAG) em gestantes e puérperas da região Norte do Brasil, com base em dados do Ministério da Saúde. Seguindo o método CRISP-DM, foram realizadas etapas de seleção, transformação e processamento dos dados, aplicação de algoritmos de aprendizado de máquina e avaliação dos modelos. Utilizaram-se os algoritmos Floresta aleatória, Regressão Logística, k-vizinhos mais próximos e *XGBoost*, com validação cruzada e avalia

O artigo de Paixão et al. (2022) apresenta o aprendizado de máquina (AM) como um

ramo da inteligência artificial voltado à construção de modelos capazes de predizer, classificar e detectar condições clínicas a partir de dados. Foi realizada uma revisão sistemática nas bases *PubMed* e *Medline*, utilizando termos relacionados a AM e cardiologia, e incluindo estudos prospectivos e retrospectivos avaliados por dois revisores independentes. A revisão identificou o uso crescente de algoritmos como redes neurais, regressão logística, floresta aleatória e SVM, especialmente na cardiologia, onde demonstraram alta precisão em diagnósticos e prognósticos. Conclui-se que o uso de AM na medicina já apresenta impacto prático, embora sua aplicação ampla ainda enfrente desafios técnicos, éticos, regulatórios e de capacitação profissional.

Embora existam estudos na literatura sobre fibrose cística, não foram encontrados trabalhos voltados à predição e classificação de variáveis específicas, isto é, modelos de aprendizado de máquina que preveem uma determinada variável de interesse, baseado na informações de outras variáveis coletadas.

1.3 Objetivos

1.3.1 Objetivo geral:

O objetivo principal deste estudo é propor técnicas de *Aprendizado de máquina* (AM) para prever a variável resposta VEF1, tanto em sua forma numérica quanto categorizada, considerando as variáveis numéricas: idade, IMC, CVF e número de consultas, além das variáveis categóricas: sexo, raça, região e genótipo.

1.3.2 Objetivos específicos:

1. Revisar na literatura os principais modelos de aprendizado de máquina utilizados em dados similares ao do estudo;
2. Caracterizar a base de dados e os principais modelos de predição e classificação, com suas métricas de avaliação (dos modelos) a serem adotados no estudo;
3. Fazer uma análise descritiva da base de dados, buscando entender melhor os dados;
4. Aplicar os modelos de aprendizado de máquina e as métricas de avaliação consideradas;
5. Comparar os modelos adotados, baseado nas métricas adotadas.

A escolha do VEF1 como variável resposta ocorreu pois o mesmo mede a quantidade de ar que uma pessoa consegue expelir com força em um segundo e o CVF, que também é uma variável importante, mede o volume total de ar que uma pessoa consegue expelir após uma inspiração máxima, contudo o CVF reflete a capacidade pulmonar total, mas não é tão sensível quanto o VEF1 para detectar a obstrução, especialmente em estágios iniciais da doença de FC.

1.4 Estrutura da dissertação

O estudo inicia pela introdução, com subseções de contextualização do tema e revisão da literatura. Após isso, segue para os materiais e métodos, com a descrição do estudo e das variáveis, população e amostra; e procedimentos usados na pesquisa. Por fim, se encontra os resultados obtidos e a conclusão.

Capítulo 2

Materiais e Métodos

2.1 Descrição do estudo

Neste estudo, foi utilizada a base de dados do Registro Brasileiro de Fibrose Cística (REBRAFC), disponibilizada pelo Grupo Brasileiro de Estudos de Fibrose Cística (GBEFC). Esse registro permite que médicos de todo o Brasil insiram, anualmente, informações sobre pacientes diagnosticados com FC, criando um banco de dados nacional atualizado sobre a condição.

2.2 População e amostra

Nesta análise foi considerada toda a população brasileira dividida nas cinco regiões, no qual o diagnóstico dos pacientes com FC vão de 1958 até 2021.

A base original contém 14.228 observações, contudo como havia muitos valores atípicos e faltantes na base, foi realizado o tratamento dos dados no estudo com base na exclusão. Dessa forma, a base final ficou com 12.338 observações. No estudo é proposto os dados brutos, sem considerar agrupamento ou redução de dimensionalidade.

2.3 Descrição das variáveis

- VEF1 (litros): Volume expiratório forçado em 1 segundo - volume de ar que o paciente foi capaz de expirar no primeiro segundo do teste de espirometria;
- CVF (litros): Capacidade vital forçada - volume máximo de ar que o paciente foi capaz de expirar durante uma expiração total no teste de espirometria;
- Consultas (quantidade);
- Idade (anos);

- IMC (kg/m^2);
- Sexo: Masculino e Feminino;
- Raça: Branco, Mestiço (pardo) e Negro;
- Região: Sul, Sudeste, Centro-Oeste, Norte e Nordeste;
- Genótipo: Sim/Não - Presença ou ausência desse procedimento, utilizado para identificação de possíveis genes patogênicos relacionados à Fibrose Cística no paciente.

2.4 Metodologia

Em *Aprendizado de máquina* (AM), há duas categorias principais de algoritmos: supervisionado e não supervisionado. Em aprendizado supervisionado, o modelo é treinado com dados rotulados, onde o objetivo é prever um rótulo ou valor específico. Já no aprendizado não supervisionado, o modelo não tem rótulos e busca identificar padrões ou agrupamentos nos dados.

O estudo tem o foco no aprendizado supervisionado, analisando três modelos de predição (com VEF1 sendo numérica): Modelos Lineares Generalizados (MLG) com resposta gama, Árvores de Regressão e Florestas Aleatórias. E três modelos de classificação (com VEF1 sendo binária): Modelos Lineares Generalizados (MLG) com resposta logística, Árvores de Regressão e Florestas Aleatórias.

Todas as análises foram realizadas no *software R* (R Core Team (2025)), versão $4.5.0$.

2.4.1 Árvores de regressão

A obtenção da árvore de regressão, segundo Izbicki and dos Santos, 2020, envolve a divisão recursiva do espaço das variáveis explicativas. Cada divisão é chamada de nó, e os resultados obtidos são conhecidos como folhas.

Para prever uma nova observação com o método, o processo se inicia no topo da árvore e verificamos a condição do primeiro nó. Se a condição for atendida, seguimos pela esquerda; caso contrário, seguimos pela direita. segue assim até chegarmos a uma folha. Se a primeira condição for atendida, a previsão será F1. Caso contrário, seguimos para a direita e encontramos outra condição. Se esta for satisfeita, a previsão será F2; se não, será F3, como mostra na Figura 1.

Com este método é fácil compreender a relação entre as variáveis explicativas e a variável resposta, ao contrário do que ocorre com a maior parte dos métodos não paramétricos.

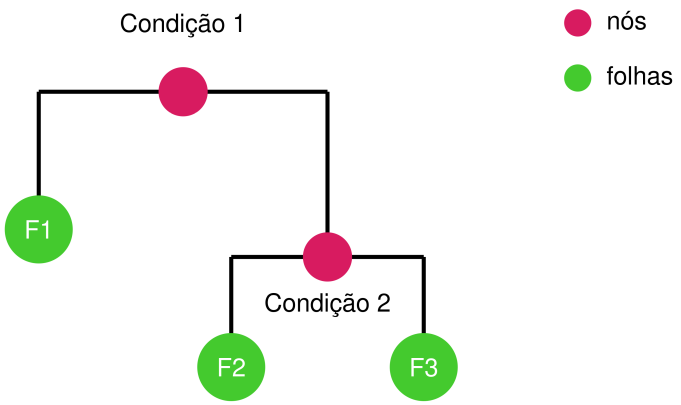


Figura 1 – Exemplo de estrutura de uma árvore de regressão.
Fonte: Izbicki and dos Santos, 2020

De maneira formal, uma árvore cria uma partição do espaço das covariáveis em regiões distintas e disjuntas: $R_1, \ldots, R_j$. A predição para a resposta Y de uma observação com covariáveis x que estão em R_k é dada por

$$g(x) = \frac{1}{|i : x_i \in R_k|} \sum_{i:x_i \in R_k} y_i. \tag{2.1}$$

Isto é, para prever o valor da resposta de x, observamos a região a qual a observação x pertence e, então, calculamos a média dos valores da variável resposta das amostras do conjunto de treinamento pertencentes àquela mesma região.

A criação da estrutura de uma árvore de regressão é feita através de duas grandes etapas: (i) a criação de uma árvore completa e complexa e (ii) a poda dessa árvore, com a finalidade de evitar o super ajuste.

No passo (i), cria-se uma árvore que leve a partições "puras", ou seja, partições nas quais os valores de Y nas observações do conjunto de treinamento em cada uma das folhas sejam homogêneos. Para isso, avalia-se o quão razoável uma dada árvore T é através de seu erro quadrático médio,

$$P(T) = \sum_R \sum_{i:x_i \in R_k} \frac{(y_i - \widehat{y_R})^2}{n},$$

em que $\widehat{y_R}$ é o valor predito para a resposta de uma observação pertencente à região R. Encontrar T que minimize $P(T)$ é inviável computacionalmente. Assim, utiliza-se uma heurística para encontrar uma árvore com erro quadrático médio baixo que consiste na criação de divisões binárias recursivas, como mostrado nas Figuras 2 a 5. Inicialmente, o algoritmo

particiona o espaço de covariáveis em duas regiões distintas (Figura 2). Para escolher essa partição, busca-se, dentre todas as covariáveis x_i e cortes t_1, aquela combinação que leva a uma partição $(R1, R2)$ com predições de menor erro quadrático:

$$\sum_{i:x_i \in R_1}^{n} (y_i - \widehat{y}_{R_1})^2 + \sum_{i:x_i \in R_2}^{n} (y_i - \widehat{y}_{R_2})^2,$$

em que $\widehat{y}_{R_k}$ é a predição fornecida para a região R_k (veja Eq. 2.1). Logo, define-se

$$R_1 = \{\mathbf{x} : x_i < t_1\} \quad \text{e} \quad R_2 = \{\mathbf{x} : x_i < t_1\},$$

em que x_i é a variável escolhida e t_1 é o corte definido.

Uma vez estabelecidas as regiões, o nó inicial da árvore é fixado. No passo seguinte busca-se particionar R_1 ou R_2 em regiões menores (Figura 3). Para escolher a nova divisão, a mesma estratégia é utilizada: dentre todas as covariáveis x_i e cortes t_2, se busca aquela combinação que leva a uma partição com menor erro quadrático. Note que agora também é necessário escolher qual região deve ser particionada: R_1 ou R_2. Assuma que R_1 foi a região escolhida, juntamente com a covariável x_j e o corte t_2. Chamemos a partição de R_1 de $\{R_{1,1}, R_{1,2}\}$ como mostra a Figura 3. Dessa forma,

$$R_{1,1} = \{\mathbf{x} : x_i < t_1, x_j < t_2\}, R_{1,2} = \{\mathbf{x} : x_i < t_1, x_j \geq t_2\}, R_2 = \{\mathbf{x} : x_i \geq t_1\}.$$

O processo segue recorrente, como nas Figuras 4 e 5, até que cheguemos a uma árvore com poucas observações em cada uma das folhas.

A árvore criada usando deste procedimento obtém bons resultados para o conjunto de treinamento, porém há grandes chances de que ocorra o super-ajuste e isso gera uma performance preditiva ruim em novas observações. Desta maneira, prossegue-se para o passo (ii), que é chamado de poda, que tem como objetivo tornar a árvore de regressão menor e menos complexa, de modo a diminuir a variância desse estimador. Nessa etapa do processo, cada nó é retirado um por vez, e nota-se como o erro de predição varia no conjunto de validação. Com base nisso, é decidido quais nós seguirão na árvore.

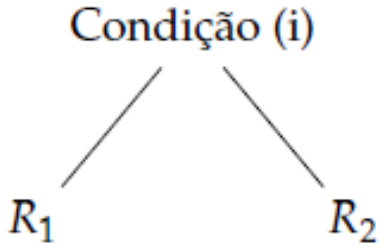


Figura 2 – Processo de crescimento de uma árvore de regressão

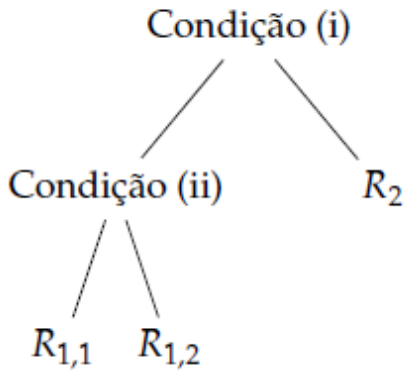


Figura 3 – Processo de crescimento de uma árvore de regressão

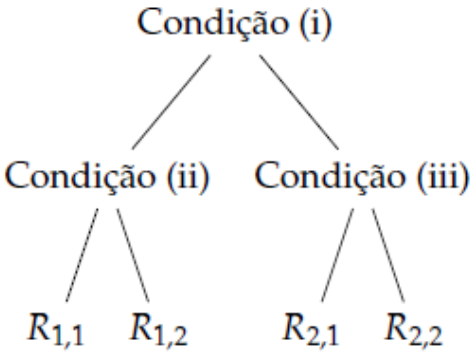


Figura 4 – Processo de crescimento de uma árvore de regressão

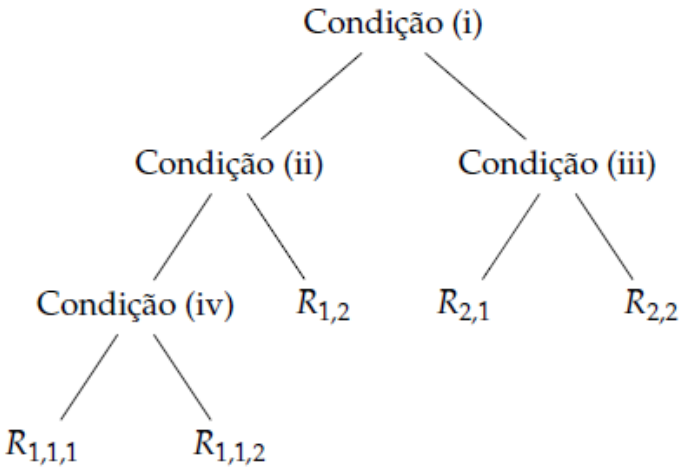


Figura 5 – Processo de crescimento de uma árvore de regressão

2.4.2 Bagging e Florestas Aleatórias

Devido as árvores de regressão, muitas vezes, apresentarem baixo poder preditivo quando comparadas aos demais estimadores, Bagging e florestas aleatórias Breiman (2001)

surgiram como métodos que contornam essa limitação combinando diversas árvores para fazer uma predição para um mesmo problema.

Como por exemplo, em um contexto de regressão, temos duas funções de predição para Y, $g_1(x)$ e $g_2(x)$. Os riscos dessas (condicionais em x, mas não na amostra de treinamento) são dados, respectivamente, por

$$\mathbb{E}\left[(Y - g_1(x))^2|x\right] \quad \text{e} \quad \mathbb{E}\left[(Y - g_2(x))^2|x\right].$$

Considerando o estimador combinado $g(x) = (g_1(x) + g_2(x))/2$, temos que

$$\mathbb{E}\left[(Y - g(x))^2|x\right] = \mathbb{V}\left[Y|x\right] + \frac{1}{4}(\mathbb{V}\left[g_1(x) + g_2(x)|x\right]) + \left(\mathbb{E}\left[Y|x\right] - \frac{\mathbb{E}\left[g_1(x)|x\right] + \mathbb{E}\left[g_2(x)|x\right]}{2}\right)^2.$$

Deste modo, se $g_1(x)$ e $g_2(x)$ são não correlacionados ($Cor(g_1(x), g_2(x)|x) = 0$), não viesados ($\mathbb{E}[g_1(x)|x] = \mathbb{E}[g_2(x)|x] = r(x)$) e têm mesma variância ($\mathbb{V}[g_1(x)|x] = \mathbb{V}[g_2(x)|x]$), temos

$$\mathbb{E}\left[Y - g(x)^2|x\right] = \mathbb{V}\left[Y|x\right] + \frac{1}{2}\mathbb{V}\left[g_i(x)|x\right] \leq \mathbb{E}\left[(Y - g_i(x))^2|x\right], \tag{2.2}$$

com $i = 1, 2$. Desta maneira, é melhor utilizar o estimador combinado $g(x)$ do que usar $g_1(x)$ ou $g_2(x)$ separadamente. Mesmo que se considere apenas dois estimadores da função de regressão nesse exemplo, a conclusão permanece válida quando se combina um número B qualquer de estimadores.

Os métodos de *bagging* e florestas aleatórias se valem dessa ideia para melhorar as predições dadas por árvores. Os dois métodos constituem em criar B árvores distintas e combinar seus resultados para melhorar o poder preditivo em relação a uma árvore individual. Para criação das B árvores distintas, o *bagging* utiliza B amostras bootstrap da base original. Para cada uma dessas amostras, cria-se uma árvore usando as técnicas descritas na Subseção 2.4.1. Contudo, assumimos na derivação da Equação 2.2 que os estimadores são não viesados. Assim, para criar árvores próximas de não-viesadas, não podamos as árvores criadas, ou seja, utilizamos apenas o passo (i) descrito na seção $2.4.1$.

Seja $g_b(x)$ a função de predição obtida segundo a $b - ésima$ árvore. A função de predição dada pelo *bagging* é dada por

$$g(x) = \frac{1}{B}\sum_{b=1}^{B} g_b(x).$$

Apesar do método *bagging* retornar preditores que não são tão fáceis de interpretar, o mesmo permite a criação de uma medida de importância para cada variável explicativa. Essa medida de importância é baseada na redução da soma de quadrados dos resíduos (SQR) 1de cada divisão. Para ilustrar o cálculo dessa quantidade, considere a situação da Figura 6, na qual dividimos o nó "pai"em dois grupos: esq (esquerda) e dir (direita).

A redução no SQR pela variável utilizada é dada por

$$RSS_{pai} - RSS_{esq} - RSS_{dir} = \sum_{i \in pai} (y_i - \bar{y}_{pai})^2 - \sum_{i \in esq} (y_i - \bar{y}_{esq})^2 - \sum_{i \in dir} (y_i - \bar{y}_{dir})^2.$$

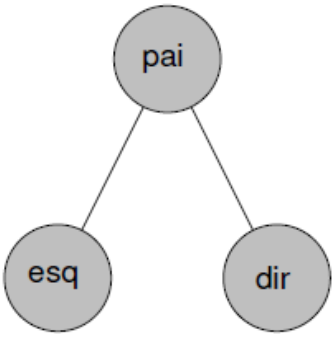


Figura 6 – Divisão de uma árvore usada para a ilustração do cálculo do SQR.

Para calcular a medida de importância de uma variável, primeiramente se calcula o total da redução da soma de quadrados dos resíduos (SQR) para todas as divisões que foram feitas com base nessa variável. Esse processo é repetido para todas as árvores e calcula-se a média dessa redução. Essa redução média é então usada como medida de importância para cada variável.

2.4.2.1 Florestas Aleatórias

Na Equação 2.2, além da suposição de que os estimadores são não viesados, assumimos que são não correlacionados. Em geral, mesmo usando amostras bootstrap para construir cada uma das árvores, as árvores construídas tendem a dar predições parecidas. Com isso, as funções $g_b(x)$ resultantes, em geral, apresentam alta correlação.

As Florestas Aleatórias tentam diminuir esta correlação modificando o mecanismo de criação das árvores para que essas se tornem diferentes umas das outras. Isto é, ao invés de escolher qual das d covariáveis será utilizada em cada um dos nós da árvore, em cada passo só é permitido que seja escolhida uma dentre as $m < d$ covariáveis. Estas m covariáveis são escolhidas aleatoriamente dentre as covariáveis originais e, a cada nó criado, um novo subconjunto de covariáveis é sorteado.

O valor de m pode ser escolhido via validação cruzada. Resultados empíricos sugerem, de forma geral, que $m \approx d/3$ leva a uma boa performance. Note-se ainda, que quando $m = d$, o método de florestas aleatórias se reduz ao *bagging*.

Essa abordagem dispõem um poder preditivo em geral muito maior que o de árvores de regressão. Normalmente, o ajuste obtido com floresta aleatória é mais suave que o ajuste obtido com a árvore de regressão, particularmente, ele se aproxima melhor da regressão real.

2.4.3 Modelos Lineares Generalizados (MLG)

Os MLGs são uma extensão do modelo de regressão linear que permite modelar variáveis dependentes que seguem distribuições diferentes da normal, como binomial, Poisson, entre outras. Eles são amplamente utilizados quando as suposições de normalidade e homocedasticidade (variância constante) não se aplicam aos dados (Nelder and Wedderburn (1972)).

Em uma amostra com n observações $(y_i, \mathbf{x_i})$ onde $\mathbf{x_i} = (x_{1i}, \ldots, x_{ki})^T$ é um vetor de covariáveis, os modelos envolvem 3 componentes, sendo eles:

i. **Componente Aleatório:** um conjunto de variáveis aleatórias independentes $Y_1, \ldots, Y_n$ provenientes de uma distribuição pertencente à família exponencial (FE) uni ou biparamétrica (binomial, normal, poisson, gama, normal inversa, entre outras).

ii. **Componente Sistemático:** as variáveis explicativas entram na forma de uma soma linear de seus efeitos da seguinte forma:

$$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_i,$$

ou na forma matricial: $\eta = X\beta$, em que X é a matriz de covariáveis, $\beta = (\beta_1, \ldots, \beta_k)^T$ é o vetor de parâmetros e $\eta = (\eta_1, \ldots, \eta_k)^T$ o preditor linear.

Os MLGs são modelos de regressão lineares nos parâmetros.

iii. **Função de Ligação:** é uma função que relaciona o componente aleatório Y ao componente sistemático η, ou seja, relaciona a média de Y ao preditor linear da seguinte forma:

$$\eta_i = g(\mu_i),$$

onde $g(.)$ é a função de ligação, monótona e duas vezes diferenciável.

A escolha da função de ligação depende do problema em particular, e principalmente do espaço paramétrico de μ da distribuição selecionada. Algumas funções de ligação mais usuais são:

- Logarítmica: $\eta = \log \mu$, para $\mu \in \mathbb{R}_+$;

- Potência: $\eta = \mu^{\lambda}$, para um determinado $\lambda \in \mathbb{R}$;
- Logística: $\eta = \log\left(\frac{\mu}{1-\mu}\right)$, para um determinado $\mu \in (0,1)$;
- Probit: $\eta = \phi^{-1}(\mu)$, em que $\phi(.)$ é a função de distribuição acumulada da distribuição normal padrão;
- Cauchit: $\eta = \tan[\pi(\mu - 0.5)]$;
- Inversa: $\eta = \frac{1}{\mu}$.

2.4.4 Modelo de regressão com resposta gama

A distribuição gama pode ser usada na modelagem de dados contínuos assimétricos positivos (ou simétricos, onde a relação variância-média seja quadrática). Sendo assim, seja Y uma variável aleatória que tem distribuição gama de média μ e coeficiente de variação $1/\sqrt{\nu}$, sua função densidade de probabilidade é dada por

$$f(y;\mu,\phi) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu}\exp\left(-\frac{\nu y}{\mu}\right)y^{-1}, y>0; \mu>0; \nu>0,$$

em que $\Gamma(\nu) = \int_0^{\infty} t^{\nu-1}e^{-t}dt$.

Para a distribuição gama, a variância de Y é dada por $\text{Var}(Y) = \phi\mu^2$, onde o parâmetro de dispersão ϕ é definido como $\phi = \nu^{-1}$. Ainda que a variância dependa da média, o coeficiente de variação permanece constante. Conforme ϕ aumenta, a distribuição de Y tende a se aproximar de uma distribuição normal com média μ e variância $\mu^2\phi$.

No cenário de MLG, pode-se considerar

$$y_i|x_i \sim gama(\mu_i, \phi)$$

$$g(\mu_i) = \beta_1 + \beta_2 x_{i1} + \ldots \beta_p x_{ip}, i = 1, \ldots, n.$$

E como alternativas de funções de ligação, temos

Função inversa (ligação canônica): $g(\mu_i) = \mu_i^{-1}$;

Função logarítmica (efeitos multiplicativos): $g(\mu_i) = \log(\mu_i)$;

Função identidade (ligação canônica): $g(\mu_i) = \mu_i$.

Para o modelo de regressão com uma resposta que segue a distribuição gama, há um parâmetro de dispersão (ϕ) que precisa ser estimado, normalmente por meio do estimador X^2 , utilizando o método dos momentos.

Com a necessidade de estimar o parâmetro de dispersão, a distribuição normal é empregada na construção de testes de hipótese e intervalos de confiança. Na análise de deviances, utiliza-se a distribuição F em vez da distribuição χ^2.

2.4.5 Modelo de regressão logístico

A regressão logística é um caso particular de MLG no qual a variável resposta Y_i é binária e segue distribuição binomial. Nesse caso, o modelo é definido pelo uso da função de ligação *logito*, sendo

$$y_i \,|\, \mathbf{x}_i \sim \text{Binomial}(m_i, \pi_i), \quad i = 1, 2, \ldots, n,$$

com observações independentes, onde m_i é o número de tentativas e π_i a probabilidade de sucesso.

O modelo de regressão logística é dado por:

$$\begin{aligned} g(\pi_i) &= \ln\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \end{aligned}$$

em que $g(\cdot)$ é a função de ligação logito e β_p os coeficientes do modelo.

De forma equivalente, o modelo pode ser escrito na escala das razões de chances:

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right),$$

ou também na escala da probabilidade (resposta):

$$\pi_i = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)}.$$

A interpretação dos parâmetros é feita com base nas razões de chances, calculadas como $OR = e^{\beta_p}$. Para modelos com múltiplas covariáveis, a interpretação do OR de uma variável é válida mantendo-se fixos os valores das demais (Kleinbaum and Klein (2010)).

2.4.6 Validação cruzada K-fold

A validação cruzada (*cross-validation*) é uma técnica utilizada para estimar o desempenho preditivo de modelos, reduzindo a variabilidade associada à divisão aleatória dos dados em conjuntos de treino e teste (Hastie et al. (2009)). Dentre as variações existentes, a validação cruzada *K-fold* é amplamente utilizada.

No procedimento K-fold, o conjunto de dados é particionado aleatoriamente em K subconjuntos aproximadamente iguais, denominados *folds*. O processo segue as seguintes etapas:

1. Dividir os dados em subconjuntos mutuamente exclusivos e de tamanho similar.
2. Para cada iteração $k = 1, 2, \ldots, K$:
   a) Utilizar o k-ésimo *fold* como conjunto de teste.
   b) Utilizar os $K - 1$ *folds* restantes como conjunto de treino.
   c) Ajustar o modelo no conjunto de treino e avaliar seu desempenho no conjunto de teste, obtendo uma ou mais métricas de interesse.
3. Calcular a média e o desvio-padrão das métricas obtidas ao longo das K iterações, resultando em uma estimativa mais estável do desempenho do modelo.

O valor de K é escolhido de acordo com o tamanho da amostra e o equilíbrio entre viés e variância. Valores típicos são $K = 5$ ou $K = 10$, que apresentam bom compromisso entre custo computacional e precisão da estimativa (Kohavi (1995)).

2.4.7 Métricas de avaliação

2.4.7.1 Erro Quadrático Médio (MSE)

O MSE é a média dos quadrados das diferenças entre os valores reais (y_i) e os valores previstos ($\hat{y}_i$):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

em que y_i é o valor real (observado), $\hat{y}_i$ é o valor predito pelo modelo e n é o número de observações.

O MSE penaliza os erros grandes, pois o método utiliza o quadrado das diferenças, na qual significa que erros mais distantes entre o valor real e o valor predito têm um peso maior.

2.4.7.2 Raiz do Erro Quadrático Médio (RMSE)

Já o RMSE é simplesmente a raiz quadrada do MSE, o que traz o erro de volta à mesma unidade dos dados originários.

$$\text{RMSE} = \sqrt{\text{MSE}}.$$

Assim como o MSE, o RMSE penaliza fortemente os erros maiores, porém ele está na mesma unidade dos dados, tornando-o mais interpretável.

2.4.7.3 Erro Absoluto Médio (MAE)

O MAE é a média das diferenças absolutas entre os valores reais e os valores previstos. A métrica é dada pela equação abaixo.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

O MAE mede o erro médio em termos absolutos e é mais robusto a outliers, porque o mesmo trata todos os erros de forma linear, sem penalizar mais os erros maiores como as métricas dadas acima.

2.4.7.4 Coeficiente de Determinação (R^2)

O R^2, dado pela fórmula abaixo, mede a proporção da variação total dos dados que é explicada pelo modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

em que $\bar{y}$ é a média dos valores reais de y_i. A primeira soma no numerador é o erro residual (diferença entre os valores reais e os previstos) e a segunda soma no denominador é a variabilidade total dos dados.

O R^2 varia entre 0 e 1, na qual $R^2 = 1$ significa que o modelo explica 100% da variabilidade nos dados, e $R^2 = 0$ significa que o modelo não explica nada da variabilidade.

Para avaliar o desempenho em modelos de classificação podemos usar como base as matrizes de confusão, como é apresentada a seguir:

Tabela 1 – Matriz de confusão

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

Fonte: Livro - Aprendizado de máquina (Izbicki and dos Santos (2020)).

Com base na Tabela 1, pode-se definir várias medidas, como:

2.4.7.5 Sensibilidade

Mede a proporção de casos positivos corretamente identificados, dado pela fórmula:

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

2.4.7.6 Especificidade

Mede a proporção de casos negativos corretamente identificados.

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

2.4.7.7 Valor preditivo positivo (VPP)

É a proporção de predições positivas que realmente são positivas. Dada por:

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

2.4.7.8 Valor preditivo negativo (VPN)

É a proporção de predições negativas que realmente são negativas.

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

2.4.7.9 Acurácia

É a medida geral de acertos do modelo.

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

2.4.7.10 Curva ROC - AUC

A curva ROC (Característica de Operação do Receptor) relaciona a $sensibilidade$ e $1 - especificidade$ para diferentes limiares de decisão. A área sob a curva (AUC) mede a capacidade geral do modelo de discriminar entre as classes:

$AUC = 0,5$: desempenho aleatório.

$AUC = 1,0$: separação perfeita.

2.4.7.11 Coeficiente Kappa de Cohen (k)

O Kappa mede o grau de concordância entre as predições do modelo e os valores reais, corrigindo pelo acaso.

$$k = \frac{p_o - p_e}{1 - p_e},$$

em que p_o é a proporção de concordância observada e p_e é a proporção de concordância esperada ao acaso e são dados por:

$$p_o = \frac{VP + VN}{FP + FN + VP + VN},$$

$$p_e = \frac{(VP + FP)(VP + FN)(FN + VN)(FP + VN)}{(FP + FN + VP + VN)^2}.$$

Capítulo 3

Resultados

Neste capítulo foi proposto e analisado inicialmente a parte descritiva dos dados e em seguida a aplicação de AM de predição e classificação considerando VEF1 como variável de interesse.

3.1 Análise descritiva

Tabela 2 – Estatísticas descritivas das variáveis quantitativas.

Variável	Média	Desvio Padrão	Mínimo	Mediana	Máximo
Consultas	$4,09$	$1,62$	$1,00$	$4,00$	$8,00$
CVF	$2,46$	$1,03$	$0,21$	$2,31$	$5,51$
Idade	$24,64$	$8,01$	$8,00$	$23,00$	$47,00$
IMC	$18,51$	$3,28$	$8,77$	$18,18$	$29,01$
VEF1	$1,88$	$0,83$	$0,17$	$1,74$	$4,28$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Nota-se na Tabela 2, que a variável resposta VEF1 apresenta uma média de $1,88$, com um valor mínimo de $0,17$ e um máximo de $4,28$. A mediana é de $1,74$, o que indica que a distribuição pode ser ligeiramente inclinada para a direita, já que a média é maior que a mediana. O desvio padrão de $0,83$ sugere uma dispersão razoável.

A variável CVF tem uma média de $2,46$, com o valor mínimo de $0,21$ e o máximo de $5,51$. O desvio padrão de $1,03$ também indica uma dispersão significativa.

Quanto a Idade, apresenta uma média de $24,64$ anos, com valores variando de 8 a 47 anos. Já o IMC dos pacientes variam de $8,77$ a 29, a média é de 19.

Por fim, os pacientes de FC tem em média de 4 consultas, com valores variando de 1 a 8. O desvio padrão de $1,62$ indicam que a maioria dos indivíduos realizou entre 2 e 4 consultas.

Na análise a seguir, encontra-se as relações de frequência das variáveis qualitativas por região.

Tabela 3 – Distribuição percentual da variável sexo por região.

Região	**Sexo – Feminino (%)**	**Sexo – Masculino (%)**
Centro-Oeste	$51,82$	$48,18$
Nordeste	$51,76$	$48,24$
Norte	$39,68$	$60,32$
Sudeste	$47,47$	$52,53$
Sul	$51,64$	$48,36$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Quanto ao distribuição percentual do sexo por região na Tabela 3, as regiões Centro-Oeste, Nordeste e Sul apresentam leve predominância de indivíduos do sexo feminino (51%). No Sudeste, nota-se um pequeno predomínio masculino $(52,5\%)$. Já o Norte se destaca com a maior proporção de homens $(60,3\%)$ e a menor de mulheres $(39,7\%)$, indicando uma desigualdade acentuada na distribuição por sexo nessa região.

Tabela 4 – Distribuição percentual da covariável genótipo por região.

Região	**Genótipo – Sim (%)**	**Genótipo – Não (%)**
Centro-Oeste	$95,15$	$4,85$
Nordeste	$85,29$	$14,71$
Norte	$86,08$	$13,92$
Sudeste	$92,68$	$7,32$
Sul	$92,54$	$7,46$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Na Tabela 4, a variável genótipo apresenta alta prevalência da categoria "Sim"em todas as regiões do país, as maiores proporções foram observadas no Centro-Oeste $(95,1\%)$, Sudeste $(92,7\%)$ e Sul $(92,5\%)$. Embora levemente inferiores, as regiões Norte $(86,1\%)$ e Nordeste $(85,3\%)$ também apresentaram valores altos.

Tabela 5 – Distribuição percentual da covariável raça por região.

Região	**Raça – Branco (%)**	**Mestiço (%)**	**Negro (%)**	**Asiático (%)**	**Índio (%)**
Centro-Oeste	$68,11$	$28,42$	$3,47$	$0,00$	$0,00$
Nordeste	$38,35$	$54,47$	$6,82$	$0,24$	$0,12$
Norte	$27,61$	$70,07$	$2,32$	$0,00$	$0,00$
Sudeste	$78,29$	$13,91$	$7,47$	$0,31$	$0,01$
Sul	$95,90$	$2,74$	$1,16$	$0,19$	$0,00$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

A distribuição racial por região da tabela 5, revela padrões distintos no Brasil. A região Sul apresenta a maior proporção de pessoas brancas $(95,9\%)$, seguido pelo Sudeste $(78,3\%)$, ambas com baixa diversidade racial. Em contraste, o Norte $(70,1\%)$ e o Nordeste $(54,5\%)$ destacam-se pela predominância de pessoas mestiças (pardas), refletindo altos níveis

de miscigenação. O Centro-Oeste apresenta perfil com $68,1\%$ de brancos e $28,4\%$ de mestiços. A raça indígena e asiática é praticamente nula em todas as regiões.

Abaixo segue o histograma da variável resposta VEF1, boxPlot das variáveis qualitativas em relação ao VEF1 nas cinco regiões e um correlograma com as variáveis quantitativas.

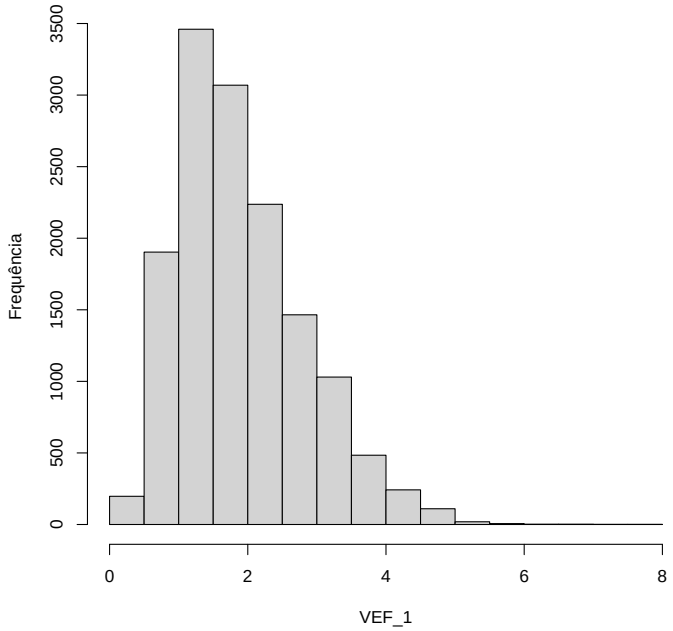


Figura 7 – Histograma da variável resposta.
Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Como podemos notar já na Tabela descritiva, o gráfico de histograma na Figura 7 nos mostra que a variável VEF1 tem em sua maioria uma distribuição assimétrica à direita.

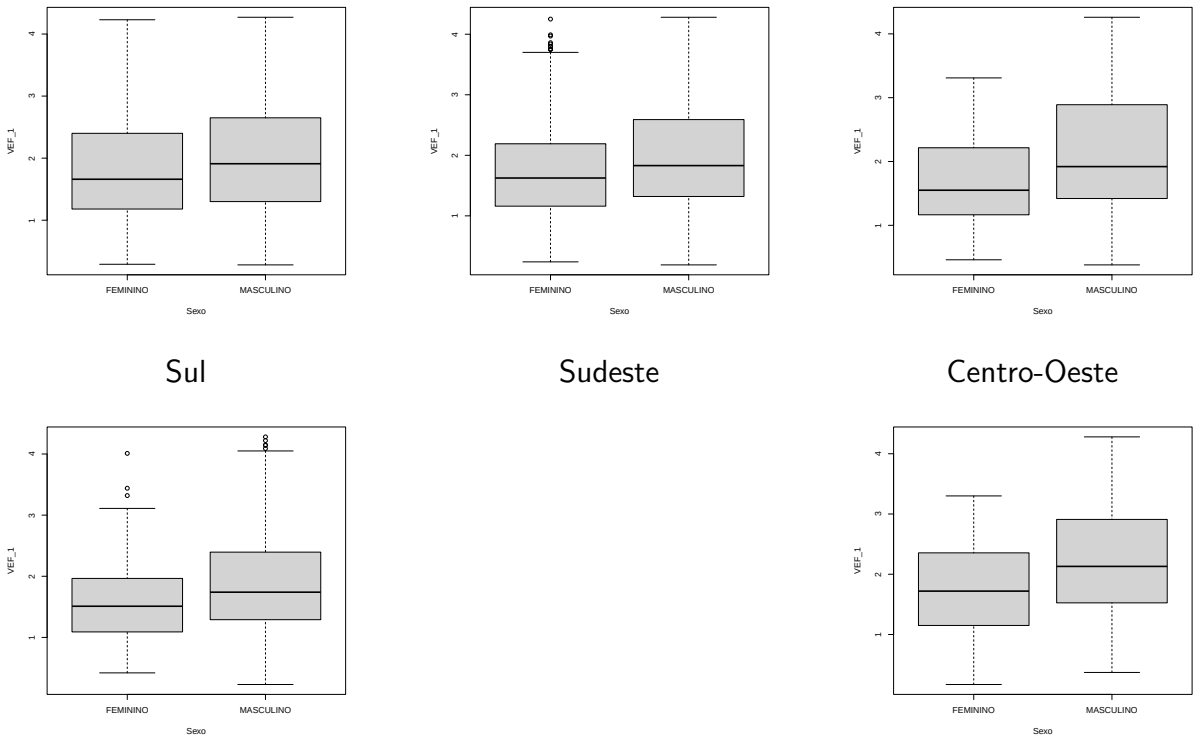


Figura 8 – Gráficos de boxplot da variável Sexo por VEF1 nas cinco regiões do Brasil.
Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Na Figura 8, que mostra gráficos de boxplot da variável VEF1 por sexo nas regiões, podemos ver que homens apresentam valores medianos de VEF1 mais altos do que as mulheres em todas as regiões. A dispersão dos dados também tende a ser maior no grupo masculino em algumas regiões, indicando maior variação individual.

Não é possível ver indícios claros de outliers extremos, mas é possível notar pequenas variações regionais nos limites inferiores e superiores dos boxplots.

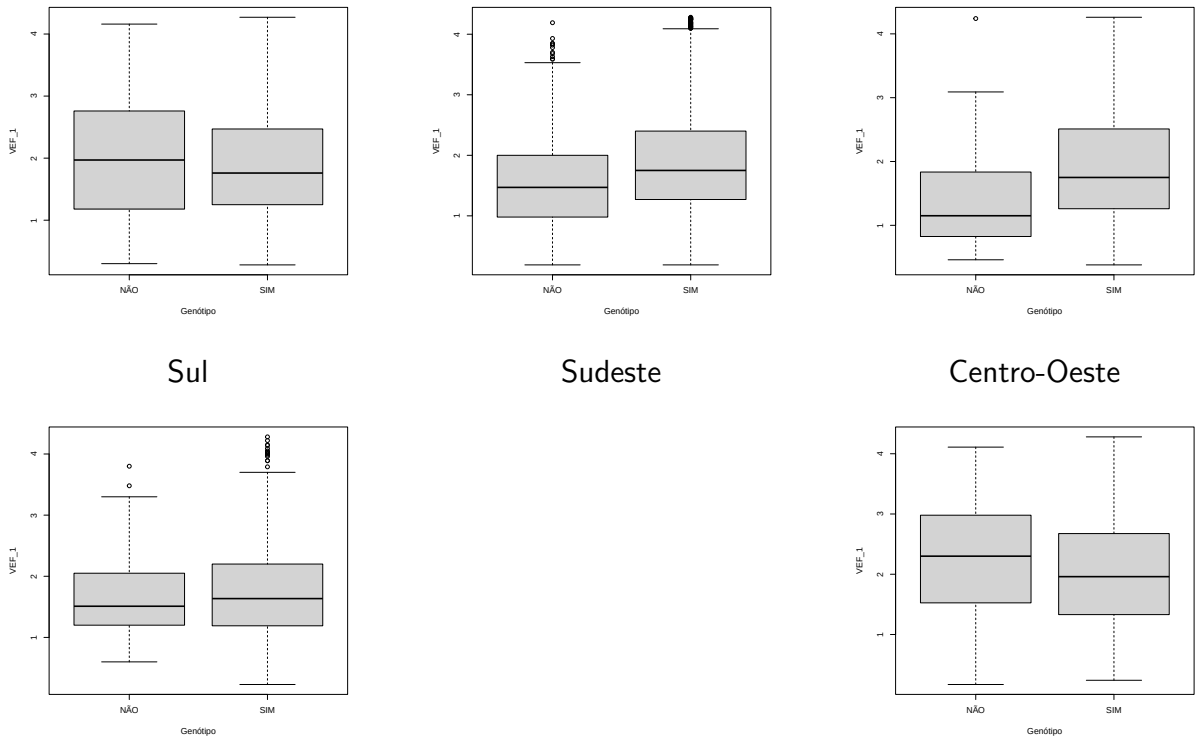


Figura 9 – Gráficos de boxplot da variável Genótipo por VEF1 nas cinco regiões do Brasil.
Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

A análise traz na Figura 9, que em três regiões, Sudeste, Centro-Oeste e no Nordeste os indivíduos com genótipo "sim" tendem a apresentar valores medianos de VEF1 ligeiramente superiores aos do grupo "não". A região Centro-Oeste, visualmente, é a região que mostra uma diferença um pouco maior entre os grupos.

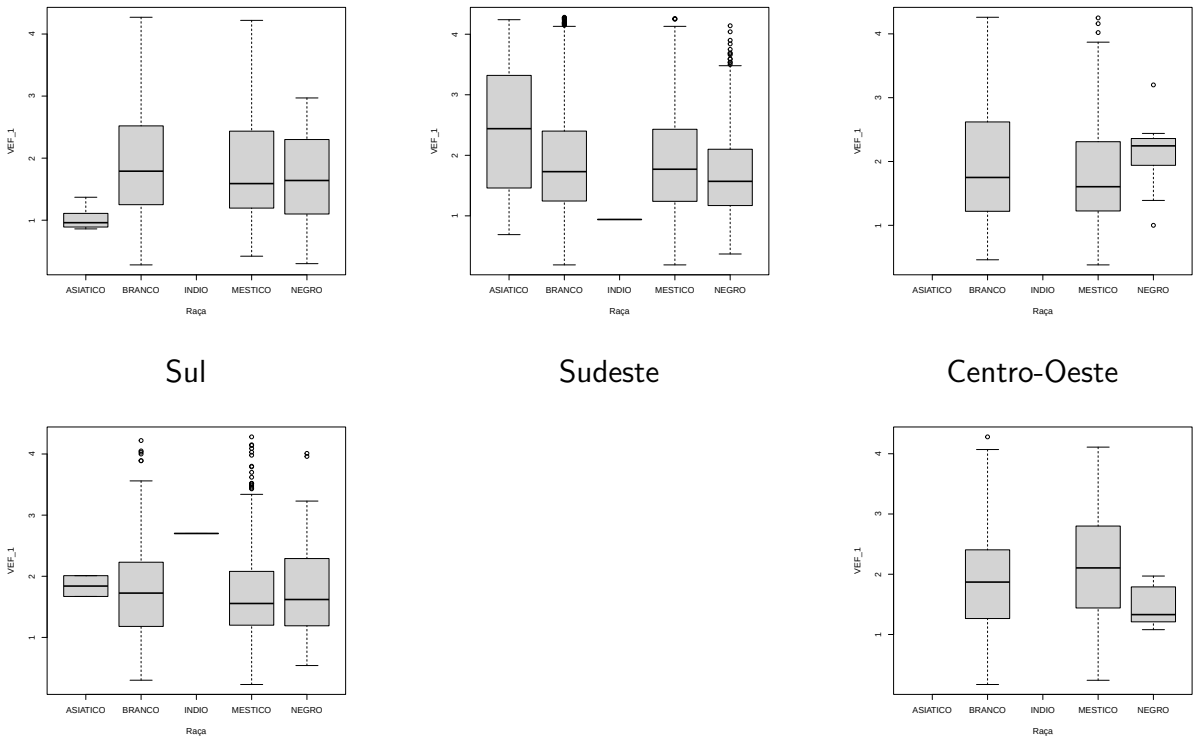


Figura 10 – Gráficos de boxplot da variável Raça por VEF1 nas cinco regiões do Brasil.
Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Quanto a análise dos gráficos de boxplot referentes ao VEF1 por raça nas cinco regiões do Brasil (Figura 10), é revelado um padrão claro e consistente na distribuição do VEF1 entre os diferentes grupos raciais asiático, branco, índio, mestiço e negro.

Na maioria das regiões analisadas, os indivíduos brancos e asiáticos apresentam os maiores valores medianos de VEF1, sugerindo melhor função pulmonar média nesses grupos, apenas na região norte se difere, com a maior mediana sendo na raça mestiça. Essa dispersão pode refletir tanto a heterogeneidade genética quanto também as desigualdades sociais que impactam a saúde respiratória da população.

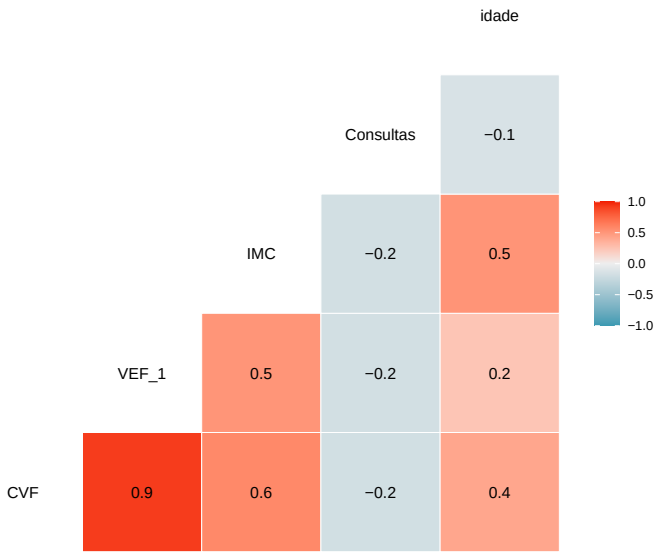


Figura 11 – Correlograma das variáveis numéricas.
Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

No correlograma da Figura 11, encontra-se a presença de correlações positivas e negativas entre as variáveis associadas a variável resposta VEF1. A mesma parece estar fortemente correlacionada com as variáveis CVF, indicando que um maior volume expiratório forçado está associado a uma maior capacidade vital forçada.

3.2 Técnicas de aprendizado de máquina

Nesta subseção, uma vez que fizemos a análise descritiva, iremos partir para a modelagem de aprendizado de máquina, atráves dos modelos de predição que é quando queremos verificar um valor exato e dos modelos de classificação, quando o objetivo é classificar em termos de grupos de VEF1.

3.2.1 Modelos de predição

O ajuste ocorreu nos modelos de árvore de regressão, floresta aleatória e modelo linear generalizado com resposta gama, que inicialmente foi realizado utilizando todas as variáveis explicativas, contudo a covariável raça não foi significativa estatisticamente e foi retirada do modelo final. A base de dados foi dividida em 70% para treino e 30% para teste, sendo aplicada validação cruzada do tipo *K-fold* para avaliar o desempenho dos modelos, com $K = 10$. As análises foram feitas no *software R* (R Core Team (2025)), utilizando os pacotes *caret* (Kuhn

et al., 2020), *rpart* (Milborrow and Milborrow, 2019), *randomForest* (RColorBrewer and Liaw, 2018) e *MASS* (Ripley et al., 2013).

Primeiramente, encontra-se o modelo de árvore de regressão expresso na Figura 12.

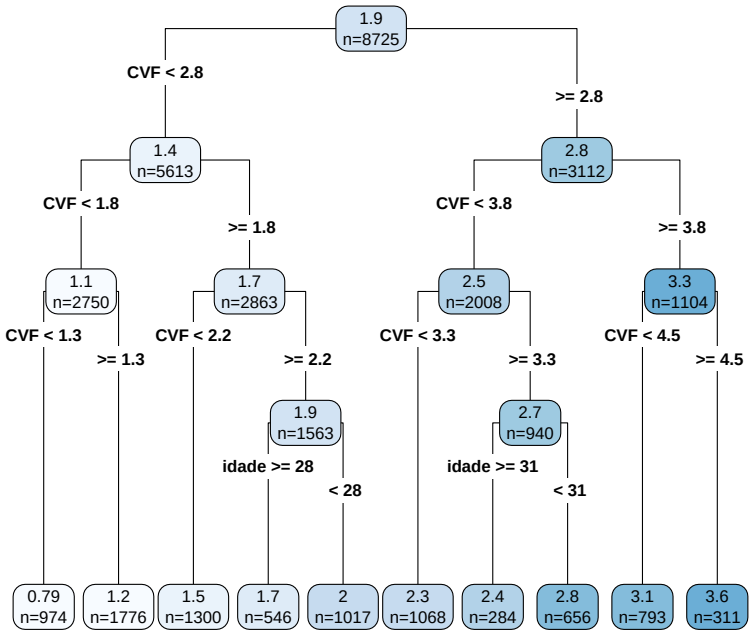
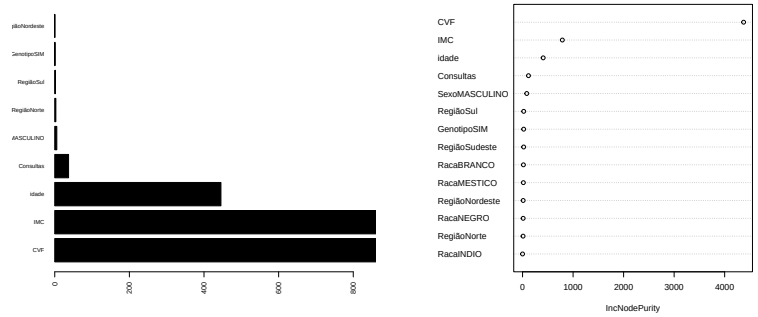


Figura 12 – Modelo de árvore de regressão.
Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Com base na Figura 12, podemos observar que a árvore de regressão destaca que CVF é a variável predominante na determinação do VEF1, com divisões principais baseadas em faixas de CVF. A covariável idade surge como uma variável relevante em níveis intermediários de CVF. Os valores de VEF1 aumentam progressivamente nos grupos com CVF mais elevado, o que faz sentido de forma clínica, pois um CVF maior geralmente está associado a melhor função pulmonar em pacientes com FC.



Árvore de regressão Floresta aleatória

Figura 13 – Importância das variáveis nos modelos analisados.

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Com os gráficos de importância (Figura 13) referentes aos modelos árvore de regressão e do floresta aleatória, que relaciona quanto cada variável contribui para aumentar a "pureza"dos nós em cada árvore do modelo, isto é, indica quanto cada variável preditora contribui para as previsões do modelo, observa-se que em ambos modelos a covariável CVF é a variável mais importante no impacto na previsão, seguido de IMC e idade, corroborando com a Figura 12.

Segue abaixo o modelo generalizado com resposta gama ajustado com todas as covariáveis consideradas significativas ao nível de 5% de significância.

Tabela 6 – Estimativas do modelo gama.

Variável	**Estimativa**	**Erro Padrão**	**Valor t**	**P-valor**
(Intercepto)	$-0,48$	$0,04$	$-13,34$	$<$ **0,001**
Sexo Masculino	$-0,06$	$0,00$	$-17,24$	$<$ **0,001**
Região Nordeste	$0,05$	$0,01$	$4,62$	**0,021**
Região Norte	$0,04$	$0,01$	$3,38$	$0,207$
Região Sudeste	$0,01$	$0,01$	$0,79$	$0,576$
Região Sul	$-0,01$	$0,01$	$-1,59$	**0,020**
Genótipo SIM	$0,04$	$0,01$	$5,80$	$<$ **0,001**
CVF	$0,44$	$0,00$	$204,07$	$<$ **0,001**
IMC	$0,01$	$0,00$	$9,30$	$<$ **0,001**
Consultas	$-0,00$	$0,00$	$-3,86$	**0,005**
idade	$-0,01$	$0,00$	$-44,63$	$<$ **0,001**

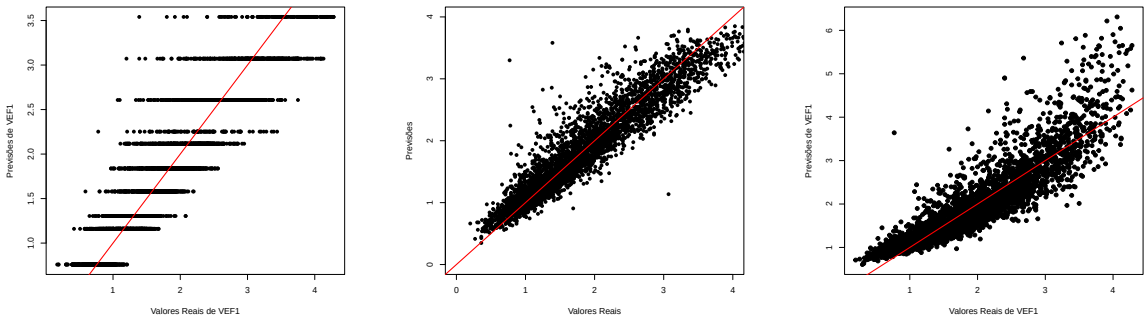
Fonte: Elaborado pelo autora, Porto Alegre/RS, 2025.

Analisando a Tabela 6 com efeitos multiplicativos, devido a função de ligação usada ser a *log*, indivíduos do sexo masculino apresentam um VEF1 aproximadamente 6% menor do que os do sexo feminino. Residir na região Nordeste está associado a uma elevação no VEF1 de cerca de 5% e na região Sul uma redução significativa de 1% em comparação a

região Centro-Oeste, enquanto que nas demais regiões não houveram diferenças significativas. A presença do genótipo tem associação a um aumento de aproximadamente $3,80\%$ no VEF1.

Quanto as variáveis contínuas, para cada unidade adicional de CVF está associada a um aumento de cerca de 55% no VEF1, reforçando sua forte relação fisiológica com a função pulmonar. O IMC apresenta um efeito positivo pequeno, já a cada ano adicional de idade reduz o VEF1 em cerca de 1%, evidenciando o decrescimento esperado com o envelhecimento. Um maior número de consultas médicas também se associa a um leve decréscimo no VEF1. Essa interpretação é válida quando as outras variáveis se mantém constantes em cada caso.

Com o propósito de analisar a qualidade dos modelos ajustados, encontra-se os gráficos de dispersão das previsões com os valores reais e após com os resíduos dos três modelos.

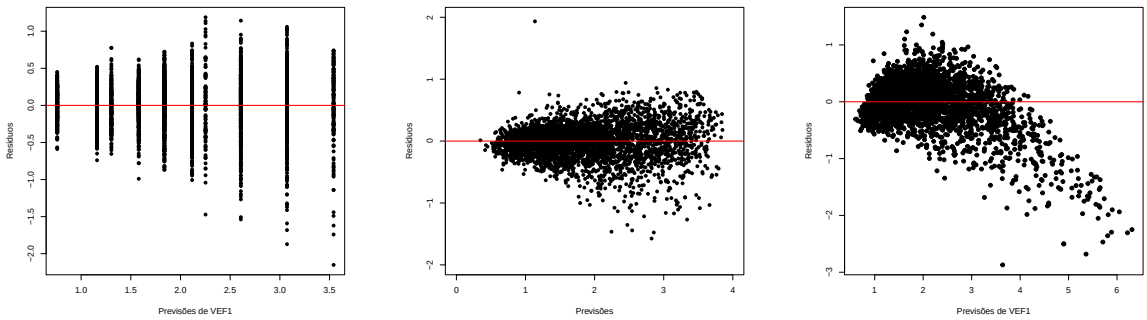


Árvore de regressão Floresta aleatória MLG com resposta gama

Figura 14 – Gráficos de dispersão das previsões vs valores reais dos modelos.

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Comparando os gráficos de dispersão entre previsões e valores reais dos três modelos (Figura 14), podemos evidenciar que a floresta aleatória apresenta melhor desempenho preditivo, com maior aderência à linha ideal e menor dispersão, especialmente nas faixas intermediárias de VEF1. A árvore de regressão mostra padrão semelhante, porém com previsões em degraus e maior variabilidade nos extremos. Já o modelo gama, embora adequado do ponto de vista teórico para variáveis positivas, apresentou viés nas faixas inferiores e menor capacidade de capturar a complexidade dos dados em comparação com os métodos baseados em árvores.



Árvore de regressão Floresta aleatória MLG com resposta gama

Figura 15 – Gráficos de dispersão das previsões vs resíduos dos modelos.

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

A Figura 15 apresenta os gráficos de dispersão entre previsões e resíduos dos três modelos analisados. Nota-se que a floresta aleatória também apresenta a melhor distribuição dos resíduos, sem padrão sistemático, mas com uma variabilidade um pouco maior no final. A árvore de regressão, embora apresente uma estrutura em degraus nos resíduos, contém a maior parte dos erros próximos de zero. Por outro lado, o modelo gama apresenta maior variabilidade e viés, com tendência de superestimar valores baixos de VEF1 e presença de heterocedasticidade, comprometendo sua capacidade de generalização em valores maiores.

Para comparar o desempenho dos três modelos propostos, foram calculadas as métricas de MAE, MSE, RMSE e o R^2, conforme apresentado na Tabela abaixo.

Tabela 7 – Métricas de comparação dos modelos.

	MAE	**MSE**	**RMSE**	R^2
Árvore de regressão	$0,27$	$0,11$	$0,34$	$0,83$
Floresta aleatória	$\mathbf{0,20}$	$\mathbf{0,08}$	$\mathbf{0,28}$	$\mathbf{0,89}$
MLG gama	$0,28$	$0,16$	$0,40$	$0,60$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Na Tabela 7, a árvore de regressão apresentou um MAE de $0,27$, um MSE de $0,11$ e um RMSE de $0,34$, relativamente maior que o modelo de floresta aleatória, sugerindo uma menor precisão na predição. Além disso, o modelo explica 83% da variabilidade dos dados. O segundo modelo, floresta aleatória, obtivemos um MAE de $0,20$, MSE de $0,08$ e RMSE de $0,28$, as menores métricas de erro entre os três modelos, logo com uma precisão superior e uma maior capacidade de capturar os padrões dos dados. Com um R^2 de $0,89$, o mais alto entre os modelos, nos mostra que este modelo explica 89% da variância da variável resposta. Por fim, o modelo gama, que embora seja estatisticamente adequado em termos de suposições de distribuição, sua capacidade preditiva foi inferior, com um MAE de $0,28$, MSE de $0,16$ e RMSE de $0,40$, é o pior modelo dentre os três. O modelo explica 60% da variabilidade dos dados.

3.2.2 Modelos de classificação

Para fins de modelagem de classificação, o VEF1 foi transformado em uma variável categórica binária, com o objetivo de prever o risco de função pulmonar reduzida em pacientes com fibrose cística. Para a modelagem da função pulmonar, utilizou-se a mediana (cenário 1) como ponto de divisão: pacientes com VEF menor ou igual à mediana foram classificados como "Baixo", e os demais como "Alto".

Os modelos ajustados foram o árvores de classificação, floresta aleatória e o modelo logístico generalizado. Nos três modelos, a base de dados foi dividida em duas partes, sendo 70% dos dados utilizados para o treinamento dos modelos e 30% para a etapa de teste, adotando a validação cruzada *K-fold* como estratégia de avaliação da performance preditiva, com $K = 10$. Os ajustes foram realizados no *software R* (R Core Team, 2025). O modelo de árvore de decisão foi ajustado com o pacote *caret* (Kuhn et al., 2020) em conjunto com o pacote *rpart* (Milborrow and Milborrow, 2019), que implementa algoritmos baseados em partições recursivas para tarefas de classificação. Já o modelo de floresta aleatória foi ajustado com o pacote *caret* (Kuhn et al., 2020) juntamente com o pacote *randomForest* (RColorBrewer and Liaw, 2018), utilizando todas as covariáveis disponíveis. No modelo logístico generalizado foi utilizando o pacote *caret* (Kuhn et al., 2020) em conjunto com o pacote *pROC* (Robin et al., 2011) e foi inicialmente ajustado com todas as covariáveis explicativas, não sendo significativas estatisticamente, as variáveis raça e número de consultas foram excluídas do modelo final.

A Figura 16 representa o modelo ajustado de árvore de classificação para predizer a função pulmonar (classificada como "Baixo"ou "Alto"com base no VEF1). O modelo indicou que a CVF foi a variável mais importante na divisão dos grupos. Indivíduos com CVF mais baixa, especialmente aqueles com idade menor e IMC reduzido, apresentaram maior probabilidade de estarem no grupo com função pulmonar comprometida. A idade e o sexo também colaboraram para a classificação em níveis mais profundos da árvore. O modelo destacou padrões clínicos relevantes e de fácil interpretação, úteis para auxiliar na triagem de pacientes com FC.

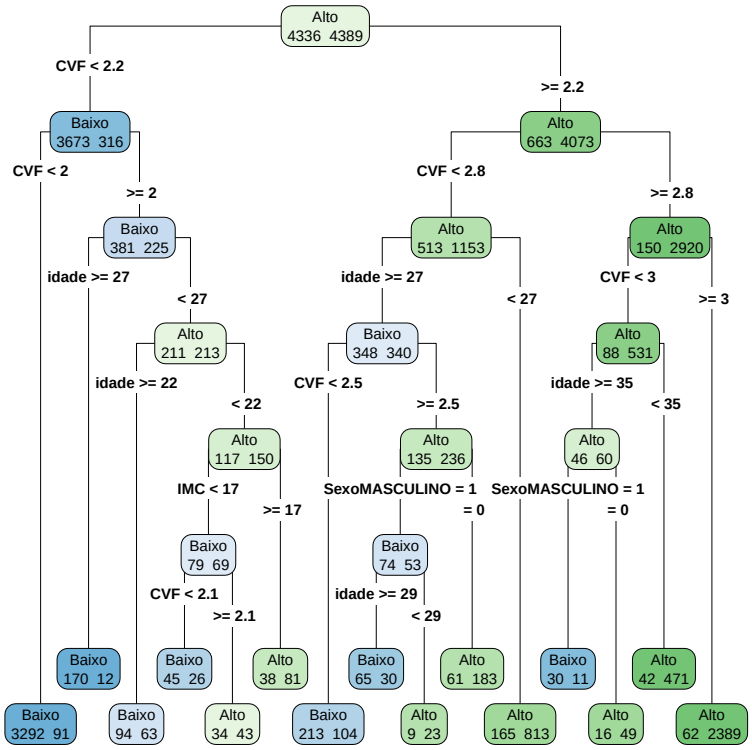


Figura 16 – Modelo de árvore de classificação.
Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

A Tabela 8 apresenta as estimativas dos coeficientes do modelo de regressão logística, juntamente com seus respectivos erros padrão, valores z e p-valores. Considerando um nível de significância de 5%, observa-se que as covariáveis sexo, região Nordeste, CVF, IMC e idade foram estatisticamente significativas no modelo final.

Tabela 8 – Estimativas do modelo de Regressão Logística.

Variável	**Estimativa**	**Erro Padrão**	**Valor z**	**P-valor**
(Intercepto)	$-9,19$	$0,37$	$-24,90$	$<$ **0,001**
Sexo Masculino	$-0,56$	$0,08$	$-6,76$	$<$ **0,001**
Região Nordeste	$0,51$	$0,23$	$2,22$	**0,026**
Região Norte	$0,44$	$0,27$	$1,60$	0,109
Região Sudeste	$-0,05$	$0,18$	$-0,29$	0,775
Região Sul	$-0,26$	$0,19$	$-1,34$	0,181
Genótipo SIM	$0,28$	$0,15$	$1,80$	0,072
CVF	$5,04$	$0,12$	$40,78$	$<$ **0,001**
IMC	$0,04$	$0,02$	$2,59$	**0,010**
Idade	$-0,14$	$0,01$	$-20,16$	$<$ **0,001**

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

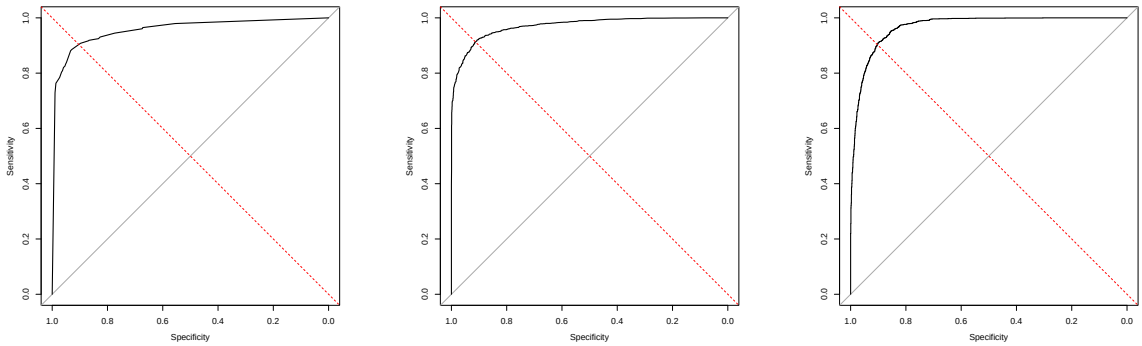
Com base nas razões de chances, analisou-se a Tabela 8 e o sexo masculino esteve associado a uma menor chance de estar no quartil de interesse "Alto" (valores acima da mediana), com uma razão de chances estimada em aproximadamente $0,57 (p < 0,0001)$. Isso implica que pacientes do sexo masculino possuem uma chance 43% menor de pertencer ao quartil superior em relação às mulheres, mantendo-se constantes as demais variáveis do modelo.

A variável região nordeste também apresentou efeitos significativos, onde teve uma razão de chances de $1,66 (p = 0,0305)$, indicando uma chance 66% maior de pertencer ao quartil superior em comparação à região de referência (Centro-Oeste).

A chance de ter VEF1 acima da mediana é cerca de 154 vezes maior para cada aumento de 1 unidade em CVF, sugerindo uma forte associação entre maior capacidade pulmonar e o desfecho analisado. Para cada ponto adicional no IMC, a chance aumentou em 4%, apesar do efeito ser relativamente pequeno, é estatisticamente relevante.

Por fim, observou-se que a variável idade apresentou um efeito negativo significativo. A cada ano adicional de idade, a chance de estar no quartil superior diminui em aproximadamente 13%, o que pode refletir um declínio natural da função associada ao desfecho com o avanço etário.

Para avaliar o desempenho dos modelos ajustados, foi realizada a curva ROC e seu respectivo AUC, que relaciona a sensibilidade e a especificidade em diferentes pontos de corte e mede a capacidade do modelo em discriminar as classes.



Árvore de regressão Floresta aleatória MLG logístico

Figura 17 – Gráficos de curva ROC dos modelo.

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

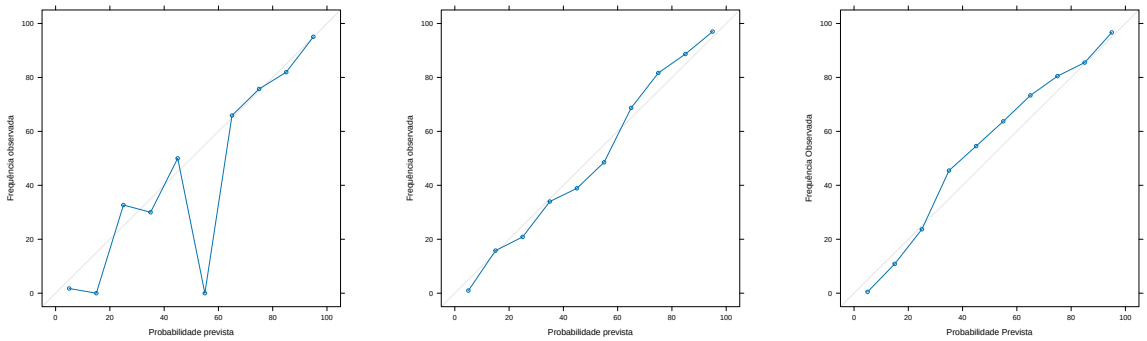
Analisando a Figura 17, pode-se notar de modo geral que todos os modelos demonstraram boa capacidade discriminativa entre as classes "Baixo" e "Alto" da função pulmonar, com áreas sob a curva (AUC) superiores a $0,96$.

O modelo de árvore de classificação obteve uma AUC de $0,96$, indicando desempenho satisfatório na separação entre os grupos. Apesar de sua simplicidade e fácil interpretação,

esse modelo apresentou leve inferioridade em relação aos demais quanto à capacidade discriminativa. A floresta aleatória destacou-se com AUC de $0,97$, sendo o modelo com melhor desempenho entre os três. O modelo logístico generalizado, também com AUC de $0,97$, apresentou desempenho semelhante ao da floresta aleatória.

Os resultados dos gráficos evidenciam que a curva ROC ficou consistentemente acima da linha de aleatoriedade nos três casos, mas embora todos os modelos tenham apresentado boa performance, a floresta aleatória e o modelo logístico generalizado se destacaram com as maiores áreas sob a curva ROC, sendo mais eficazes na separação entre os níveis de função pulmonar.

Com a finalidade de avaliar se as probabilidades previstas pelos modelos refletem, de fato, a frequência observada dos eventos, observou-se o gráfico de calibração dos três modelos.



Árvore de regressão Floresta aleatória MLG logístico

Figura 18 – Gráficos de calibração dos modelos.

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Os gráficos de calibração da Figura 18, mostram que a floresta aleatória apresenta a melhor adequação probabilística, com curva próxima da ideal e excelente correspondência entre probabilidades previstas e frequências observadas. O MLG logístico também demonstrou bom desempenho calibrado, embora com um pouco de subestimação intermediária. Já o modelo de árvore de classificação obteve maior afastamento da linha de identidade, apresentando maiores oscilações, indicando baixa calibração probabilística.

Para fins de comparação, foram definidos outros dois cenários além da classificação pela mediana, para notar como VEF1 se comporta quando o ponto de corte é pequeno, mediano e um pouco mais alto. No cenário 2, a classificação foi feita a partir do primeiro quartil (Q1), sendo os indivíduos com VEF menor ou igual a que Q1 alocados no grupo "Baixo", e os demais no grupo "Alto". Já no cenário 3, o ponto de corte adotado foi o terceiro quartil (Q3), com pacientes com VEF maior que Q3 compondo o grupo "Alto", e os demais sendo classificados como "Baixo". Em todos os cenários, nenhum paciente foi excluído da análise, o que permitiu explorar diferentes níveis de severidade da função pulmonar e sua capacidade de discriminação pelos modelos preditivos.

Nas Tabelas abaixo se encontra as métricas de acurácia, sensibilidade, especificidade, VPP, VPN, coeficiente de Kappa e AUC de cada modelo e cenário, com o propósito de comparar os modelos.

Tabela 9 – Cenário 1: Métricas de desempenho dos modelos de classificação.

Métrica	**Árvore de classificação**	**Floresta aleatória**	**Regressão logística**
Acurácia	$\mathbf{0,91}$	$\mathbf{0,91}$	$0,90$
Sensibilidade	$0,88$	$0,89$	$\mathbf{0,91}$
Especificidade	$\mathbf{0,93}$	$0,92$	$0,89$
VPP	$\mathbf{0,92}$	$0,90$	$0,89$
VPN	$0,89$	$0,89$	$\mathbf{0,91}$
Kappa	$0,81$	$\mathbf{0,82}$	$0,80$
AUC	$0,96$	$\mathbf{0,97}$	$\mathbf{0,97}$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Tabela 10 – Cenário 2: Métricas de desempenho dos modelos de classificação.

Métrica	**Árvore de classificação**	**Floresta aleatória**	**Regressão logística**
Acurácia	$0,91$	$0,92$	$0,91$
Sensibilidade	$0,76$	$0,79$	$0,80$
Especificidade	$0,96$	$0,96$	$0,94$
VPP	$0,87$	$0,87$	$0,82$
VPN	$0,92$	$0,93$	$0,93$
Kappa	$0,75$	$0,77$	$0,75$
AUC	$0,91$	$0,96$	$0,96$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Tabela 11 – Cenário 3: Métricas de desempenho dos modelos de classificação.

Métrica	**Árvore de classificação**	**Floresta aleatória**	**Regressão logística**
Acurácia	$0,92$	$0,93$	$0,92$
Sensibilidade	$0,95$	$0,95$	$0,96$
Especificidade	$0,83$	$0,85$	$0,80$
VPP	$0,94$	$0,95$	$0,93$
VPN	$0,87$	$0,87$	$0,87$
Kappa	$0,80$	$0,81$	$0,78$
AUC	$0,94$	$0,98$	$0,97$

Fonte: Elaborado pela autora, Porto Alegre/RS, 2025.

Analisando o Cenário 1 (mediana) da Tabela 9, observa-se que o modelo de floresta aleatória apresentou o melhor desempenho considerando as métricas globais, com os maiores valores de acurácia $(0,91)$, coeficiente Kappa $(0,82)$ e AUC $(0,97)$. Esses indicadores sugerem que este modelo foi o mais eficaz em classificar corretamente os indivíduos, tanto do grupo "Alto" quanto "Baixo", além de apresentar excelente capacidade discriminativa e alta concordância com os valores reais.

A regressão logística também obteve um desempenho robusto, destacando-se especialmente na sensibilidade $(0,91)$ e VPN $(0,91)$. Isso indica maior capacidade do modelo em identificar corretamente os casos positivos, minimizando os falsos negativos. Sua AUC foi igualmente elevada $(0,97)$, evidenciando excelente poder discriminativo. Além disso, trata-se de um modelo de fácil interpretação, o que favorece sua aplicação quando há interesse na compreensão dos efeitos individuais das variáveis preditoras sobre a probabilidade do desfecho.

Já a árvore de decisão teve um desempenho levemente inferior em todas as métricas. Teve boa acurácia $(0,91)$ e especificidade $(0,93)$, mas menor sensibilidade, o que pode indicar pior equilíbrio entre classes.

Realizando a comparação entre os três cenários, em relação ao cenário com base na mediana, o cenário 2 (Q1) da Tabela 10 apresentou desempenho inferior, especialmente na sensibilidade, indicando menor capacidade de identificar corretamente pacientes com VEF1 reduzido, apesar da alta especificidade. O cenário 3 (Q3) na Tabela 11 superou levemente o cenário da mediana em quase todas as métricas, com maior acurácia e sensibilidade em todos os modelos, demonstrando melhor desempenho na identificação de pacientes com função pulmonar preservada.

Com isso, conclui-se que, embora todos os modelos tenham apresentado bons desempenhos, a floresta aleatória se destacou como o mais eficiente em todos os cenários, sendo indicada quando o objetivo principal é a precisão preditiva. Por outro lado, a regressão logística, que se manteve estatisticamente equivalente à floresta aleatória em termos de AUC, se mostra vantajosa quando é para identificar o maior número possível de positivos.

Capítulo 4

Conclusão

Na análise inicial da amostra pode-se notar padrões clínicos e demográficos importantes. Observou-se predominância de pacientes com função pulmonar preservada, mas com variação considerável nos valores de VEF1, CVF e IMC, indicando heterogeneidade no perfil funcional dos pacientes. A distribuição por sexo e idade mostraram maior comprometimento pulmonar em pacientes mais velhos e do sexo masculino, achados que se alinham à literatura sobre FC, estudos como Corey and Farewell (1996) e Izbicki and dos Santos (2020) evidenciam este fator. A estratificação por regiões apontou o destaque para a região Nordeste, associada a melhores valores de função pulmonar em alguns modelos, contudo devido ao tamanho amostral de cada grupo, essas diferenças devam ser interpretadas com cautela.

Já com as análises realizadas para predição e classificação da função pulmonar em pacientes com a doença de fibrose cística, notou-se que todos os métodos empregados (árvores de regressão e classificação, florestas aleatórias e modelos lineares generalizados gama e logístico), apresentaram um desempenho satisfatório, com variações relevantes quanto a capacidade discriminativa e interpretabilidade.

Nos modelos de predição, a floresta aleatória foi o método com melhor capacidade preditiva, obtendo menores valores de erro (MAE, MSE e RMSE) e maior explicabilidade do modelo, com $R^2 = 0{,}89$. Embora o modelo gama apresente fundamentação estatística adequada para variáveis positivas, seu desempenho foi inferior aos modelos baseados em árvores. A árvore de regressão apresentou desempenho bom, sendo mais simples de interpretar, porém menos precisa que a floresta aleatória.

Nos modelos de classificação, considerando o cenário base da mediana, tanto a floresta aleatória quanto o MLG logístico obtiveram áreas sob a curva ROC elevadas ($AUC = 0{,}97$), indicando excelente capacidade discriminativa. A floresta aleatória apresentou leve superioridade em acurácia e coeficiente Kappa, refletindo melhor um desempenho global. A regressão logística, por sua vez, destacou-se pela alta sensibilidade, o que a torna mais apropriada em contextos nos quais é prioritário identificar corretamente o maior número de casos positivos, mantendo boa interpretabilidade dos coeficientes.

A avaliação dos diferentes cenários de corte, como Q1, mediana e Q3, mostrou que a performance dos modelos se mantém consistente, contudo com ligeira vantagem do cenário Q3 na acurácia e sensibilidade, especialmente para identificação de pacientes com função pulmonar preservada.

Portanto, considerando simultaneamente a capacidade preditiva, a discriminação entre classes, estabilidade de resultados e a aplicação clínica, o modelo de floresta aleatória se apresenta como o modelo mais robusto e eficaz para este conjunto de dados. Este método demonstrou desempenho superior tanto em modelos de predição quanto em de classificação, adaptando-se bem a diferentes pontos de corte e mantendo alta precisão. No entanto, a regressão logística pode ser recomendada quando a prioridade for a interpretabilidade dos efeitos das variáveis e a maximização da sensibilidade.

Dessa forma, o presente estudo contribui de forma relevante para a área da saúde, especialmente no contexto da fibrose cística, ao oferecer evidências sólidas sobre a utilização de modelos preditivos e de classificação na avaliação da função pulmonar. A identificação de que a floresta aleatória apresenta maior precisão e robustez fornece subsídios para integrar sistemas eletrônicos de saúde, facilitando o monitoramento longitudinal e a personalização do tratamento, otimizando recursos e priorizando atendimentos conforme a gravidade estimada da doença. Além disso, ao destacar o desempenho da regressão logística em termos de interpretabilidade e sensibilidade, o trabalho aponta caminhos para aplicações clínicas voltadas à identificação criteriosa de pacientes com maior risco, favorecendo decisões terapêuticas mais assertivas. Esses achados podem orientar medidas futuras, como o desenvolvimento de sistemas de apoio à decisão clínica e protocolos de acompanhamento baseados em análise de dados, contribuindo para otimizar recursos e melhorar o prognóstico dos pacientes.

Capítulo 5

Propostas futuras

Como propostas futuras, podemos listar:

- Aplicação de técnicas multivariada de agrupamento (clusterização) para identificar grupos homogêneos de pacientes antes da utilização de modelos preditivos;
  - Agrupamento hierárquico e não hierárquico;
  - Usar grupo (ou cluster) como uma nova variável categórica nos modelos preditivos;
  - Personalização de modelos preditivos para cada grupo.
- Aplicação de técnicas multivariadas de componentes principais (PCA) e/ou Análise Fatorial, para as variáveis quantitativas;
  - Utilizar os componentes principais (e/ou os fatores) como entradas (features) nos modelos preditivos;
  - Reduzir o número de variáveis para modelos mais simples;
- Comparar os modelos preditivos obtidos com e sem o uso de técnicas multivariadas, mostrando vantagens e desvantagens em cada caso.

Referências

Andersen, D. H. (1938). Cystic fibrosis of the pancreas and its relation to celiac disease: a clinical and pathologic study. *American journal of Diseases of Children*, 56(2):344–399.

Bear, C. E., Li, C., Kartner, N., Bridges, R. J., Jensen, T. J., Ramjeesingh, M., and Riordan, J. R. (1992). Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (cftr). *Cell*, 68(4):809–818.

Bittencourt, J. A. S., Sousa Junior, C. M., Santana, E. E. C., Moraes, Y. A. C. d., Carneiro, E. C. R. d. L., Fontes, A. J. C., Chagas, L. A. d., Melo, N. A. C., Pereira, C. L., Penha, M. C., et al. (2024). Predição de síndrome metabólica e seus fatores de risco associados em pacientes com doença renal crônica utilizando técnicas de machine learning. *Brazilian Journal of Nephrology*, 46:e20230135.

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

Corey, M. and Farewell, V. (1996). Determinants of mortality from cystic fibrosis in canada, 1970–1989. *American journal of epidemiology*, 143(10):1007–1017.

da Silva Bezerra, J. H. and de Almeida, F. M. (2024). Desenvolvimento de modelos preditivos com machine learning-análise de dados para saúde de gestantes e puérperas. *InterSciencePlace*, 19.

de Cássia Firmida, M., Marques, B. L., and da Costa, C. H. (2011). Fisiopatologia e manifestações clínicas da fibrose cística. *Revista Hospital Universitário Pedro Ernesto (TÍTULO NÃO-CORRENTE)*, 10(4).

de Faria, E. J. (2007). *Investigação da associação entre os polimorfismos dos genes: MBL2, TGF-B1 e CD14 com a gravidade do quadro pulmonar na fibrose cistica*. PhD thesis, [sn].

Di Sant'Agnese, P. A., Darling, R. C., Perera, G. A., and Shea, E. (1953). Abnormal electrolyte composition of sweat in cystic fibrosis of the pancreas: clinical significance and relationship to the disease. *Pediatrics*, 12(5):549–563.

Farber, S. (1944). Pancreatic function and disease in early life v. pathologic changes associated with pancreatic insufficiency in early life. *Archives of Pathology*, 37(4):0238–0250.

Furgeri, D. T., Correia, C. A. d. A., Ribeiro, A., Ribeiro, J., and Bertuzzo, C. (2006). Haplótipos da região gênica cftr em núcleos familiares de pacientes com fibrose cística. In *Anais do 1º Congresso Brasileiro de Fibrose Cística*, pages 27–30.

Gibson, L. E. and Cooke, R. E. (1959). A test for concentration of electrolytes in sweat in cystic fibrosis of the pancreas utilizing pilocarpine by iontophoresis. *Pediatrics*, 23(3):545–549.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

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

Kleinbaum, D. G. and Klein, M. (2010). *Logistic Regression: A Self-Learning Text*. Springer, New York, 3 edition.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R. C., et al. (2020). Package 'caret'. *The R Journal*, 223(7):48.

Lemos, A. C. M., Matos, E., Franco, R., Santana, P., and Santana, M. A. (2004). Fibrose cística em adultos: aspectos clínicos e espirométricos. *Jornal Brasileiro de Pneumologia*, 30:9–13.

Milborrow, S. and Milborrow, M. S. (2019). Package 'rpart. plot'. *Plot'rpart'Models: An Enhanced Version of'plot. rpart*.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384.

Paixão, G. M. d. M., Santos, B. C., Araujo, R. M. d., Ribeiro, M. H., Moraes, J. L. d., and Ribeiro, A. L. (2022). Machine learning na medicina: revisão e aplicabilidade. *Arquivos Brasileiros de Cardiologia*, 118(1):95–102.

Quinton, P. M. (1983). Chloride impermeability in cystic fibrosis. *Nature*, 301(5899):421–422.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RColorBrewer, S. and Liaw, M. A. (2018). Package 'randomforest'. *University of California, Berkeley: Berkeley, CA, USA*.

Reis, F. J. and Damaceno, N. (1998). Fibrose cística. *J Pediatr (Rio J)*, 74(Supl 1):S76–S94.

Ribeiro, J. D., Ribeiro, M. Â. G. d. O., and Ribeiro, A. F. (2002). Controvérsias na fibrose cística: do pediatra ao especialista. *Jornal de pediatria*, 78:171–186.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., and Ripley, M. B. (2013). Package 'mass'. *Cran r*, 538(113-120):822.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):77.

Rosa, F. R., Dias, F. G., Nobre, L. N., and Morais, H. A. (2008). Fibrose cística: uma abordagem clínica e nutricional. *Revista de nutrição*, 21:725–737.

Wainwright, B. J., Scambler, P. J., Schmidtke, J., Watson, E. A., Law, H.-Y., Farrall, M., Cooke, H. J., Eiberg, H., and Williamson, R. (1985). Localization of cystic fibrosis locus to human chromosome 7cen–q22. *Nature*, 318(6044):384–385.