UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

ALBERT ALVES DOS SANTOS

MODELAGEM DO CRESCIMENTO DE FRUTOS DA GOIABEIRA "PEDRO SATO": ANÁLISE MULTIRESPOSTA COM REGRESSÃO NÃO LINEAR E ALGORITMOS DE APRENDIZADO DE MÁQUINA

CAMPINA GRANDE - PB
2024

ALBERT ALVES DOS SANTOS

MODELAGEM DO CRESCIMENTO DE FRUTOS DA GOIABEIRA "PEDRO SATO": ANÁLISE MULTIRESPOSTA COM REGRESSÃO NÃO LINEAR E ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Área de concentração: Estatística

Orientador(a): Profa. Dra. Ana Patrícia Bastos Peixoto de Oliveira
Coorientador(a): Profa. Ma. Débora de Sousa Cordeiro

CAMPINA GRANDE - PB
2024

S237m Santos, Albert Alves dos.

Modelagem do crescimento de frutos da goiabeira “Pedro Sato” [manuscrito] : análise Multiresposta com Regressão não linear e algoritmos de Aprendizado de Máquina / Albert Alves dos Santos. - 2024.

51 f. : il. color.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2024.

"Orientação : Prof. Dra. Ana Patricia Bastos Peixoto de Oliveira, Departamento de Estatística - CCT".

"Coorientação: Prof. Ma. Débora de Sousa Cordeiro, Departamento de Estatística - CCT".

1. Modelos de crescimento. 2. Random forest. 3. Agricultura de precisão. 4. Manejo agrícola. I. Título

21. ed. CDD 634.3

Elaborada por Lêda Cristina Diniz Andrade - CRB - 15/1032 BC

ALBERT ALVES DOS SANTOS

MODELAGEM DO CRESCIMENTO DE FRUTOS DA GOIABEIRA "PEDRO SATO": ANÁLISE MULTIRESPOSTA COM REGRESSÃO NÃO LINEAR E ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Área de concentração: Estatística

Aprovado em: 21/ 11/ 2024.

BANCA EXAMINADORA

Profa. Dra. Ana Patricia Bastos Peixoto de Oliveira (Orientador)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Ricardo Alves de Olinda (Examinador)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Oseas Machado Gomes (Examinador)
Universidade Estadual da Paraíba (UEPB)

A Deus, pela inspiração em minha jornada. À minha família, pelo apoio incondicional, dedico.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus, que com sua graça e bondade me sustentou ao longo desta jornada, dando-me força e sabedoria para enfrentar todos os desafios.

Agradeço imensamente à minha esposa, Luana, e ao meu filho, Raul, por estarem sempre ao meu lado, oferecendo apoio, compreensão e amor incondicionais. Vocês são meu alicerce e fonte de inspiração. Cada dia ao lado de vocês me lembra o propósito de continuar evoluindo e buscar ser uma versão melhor de mim mesmo. Luana, seu carinho e incentivo foram essenciais para que eu seguisse em frente nos momentos de dúvida e cansaço, e Raul, sua alegria e curiosidade me encheram de energia e renovaram minha motivação.

Aos meus pais, expresso minha eterna gratidão por tudo o que me proporcionaram. Todo o esforço e dedicação de vocês tornaram possível a realização deste sonho, e a vocês devo a base de tudo que sou e busco ser.

Minha sincera gratidão aos amigos que compartilham essa caminhada acadêmica comigo, Luana, Matheus, Ana Beatriz, Ramon e Amauri. Foram inúmeras as horas de conversas, estudos em grupo e desabafos que juntos dividimos. A amizade e parceria de vocês tornaram os desafios mais leves e os momentos de conquista ainda mais especiais. Cada troca de ideia, cada conselho e cada gesto de apoio foram fundamentais para que eu chegasse até aqui.

Minha gratidão especial à minha orientadora, Prof$^{\underline{a}}$. Dra Ana Patricia Bastos Peixoto de Oliveira, e à minha coorientadora, Prof$^{\underline{a}}$. Ma. Débora de Sousa Cordeiro, por todo o empenho, paciência e sabedoria compartilhada. Suas orientações e suporte foram fundamentais para a realização deste trabalho. Obrigado por acreditarem no meu potencial e por estarem sempre presentes, seja para corrigir um detalhe, esclarecer uma dúvida ou oferecer palavras de incentivo.

Agradeço também aos professores que compuseram a banca examinadora do meu TCC, o Prof. Dr. Ricardo Alves de Olinda e o Prof. Dr. Oseas Machado Gomes, pelas contribuições valiosas e pelo olhar crítico que tanto enriqueceram este trabalho. Suas sugestões e questionamentos foram fundamentais para que eu pudesse aprimorar o que desenvolvi e levar minha pesquisa a um novo nível.

Aos professores do Departamento de Estatística da Universidade Estadual da Paraíba (UEPB), minha gratidão por todo o conhecimento transmitido, que serviu como base essencial para minha formação.

E, por fim, a todos que, de alguma forma, direta ou indiretamente, contribuíram para a realização deste trabalho: aos amigos, familiares, colegas e até àqueles que, com um simples gesto de apoio, fizeram diferença. Este TCC é fruto não apenas do meu esforço, mas do suporte e carinho de todos que, em algum momento, acreditaram e me apoiaram.

A todos vocês, meu mais sincero e profundo agradecimento.

RESUMO

Este trabalho tem como objetivo modelar o crescimento de frutos da goiabeira "Pedro Sato" utilizando técnicas de Regressão não linear multiresposta e algoritmos de Aprendizado de Máquina. Foram analisadas o peso e o volume dos frutos em função do comprimento longitudinal, aplicando modelos clássicos como Logístico, *Von Bertalanffy* e *Richards*, além dos algoritmos Árvore de Decisão, *Random Forest* e Máquina de Vetores de Suporte (SVM). A avaliação dos modelos foi realizada com base nos Critérios de Informação de Akaike (AIC), o Bayesiano (BIC) e a Raiz do Erro Quadrático Médio (RMSE), permitindo a identificação do modelo mais preciso para os dados experimentais. Os resultados demonstraram que o *Random Forest* foi o modelo com melhor desempenho, apresentando menor erro residual e maior capacidade de capturar interações complexas nos dados, enquanto o modelo de *Von Bertalanffy* destacou-se na descrição do padrão global de crescimento dos frutos. A aplicação dos modelos de Aprendizado de Máquina, especialmente o *Random Forest*, mostrou-se eficaz para prever o desenvolvimento dos frutos, sendo útil para o manejo agrícola e a otimização da colheita. Este estudo contribui para o avanço da modelagem de dados biológicos, demonstrando o potencial dos modelos de Aprendizado de Máquina em complementar as abordagens tradicionais em estudos de crescimento vegetal.

Palavras-chave: modelos de crescimento; *random forest*; agricultura de precisão; manejo agrícola.

ABSTRACT

This study aims to model the growth of "Pedro Sato" guava fruits using nonlinear multivariate Regression techniques and Machine Learning algorithms. The weight and volume of fruits were analyzed as a function of longitudinal length, applying classic models such as Logistic, Von Bertalanffy and Richards, in addition to the Decision Tree, Random Forest and Suport Vetor Machine (SVM) algorithms. Model evaluation was based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Square Error (RMSE) criteria, allowing the identification of the most accurate model for the experimental data. Results showed that Random Forest achieved the best performance, with the lowest residual error and the highest capability to capture complex data interactions, while the Von Bertalanffy model excelled in describing the overall growth pattern of the fruits. The application of Machine Learning models, especially Random Forest, proved effective for predicting fruit development, making it useful for agricultural management and harvest optimization. This study contributes to the advancement of biological data modeling by demonstrating the potential of Machine Learning models to complement traditional approaches in plant growth studies.

Keywords: growth models; random forest; precision agriculture; agricultural management.

SUMÁRIO

Página

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A goiabeira, pertencente à família Mirtaceae e ao gênero *Psidium*, é uma planta nativa da América Tropical, reconhecida por sua capacidade de produzir frutos carnosos conhecidos como goiabas. O crescimento desses frutos é um processo complexo que envolve várias fases, incluindo uma intensa divisão celular, seguida por um período de expansão e, finalmente, a maturação fisiológica, em que a goiaba atinge seu estádio ideal para consumo (Zeviani; Ribeiro Júnior; Bonat, 2013). Esses processos são influenciados por diversos fatores, como condições climáticas, práticas de manejo agrícola e variáveis genéticas, o que torna fundamental a modelagem estatística para compreender as dinâmicas de crescimento e produtividade.

A modelagem estatística, especialmente os modelos de Regressão não linear multiresposta, se apresentam como uma abordagem robusta para analisar simultaneamente múltiplas variáveis dependentes. Essa técnica é particularmente útil na agricultura, no qual diferentes características dos frutos, como peso, diâmetro e volume, podem ser afetadas por um conjunto comum de variáveis independentes. O uso de modelos multiresposta permite explorar as inter-relações entre essas variáveis e identificar padrões de crescimento que não seriam evidentes por meio de análises univariadas (Bates; Watts, 1988).

Neste trabalho, busca-se aplicar a Regressão Não Linear Multiresposta para modelar simultaneamente múltiplas variáveis associadas ao crescimento dos frutos da goiabeira "Pedro Sato" (*Psidium guajava* L.), com foco no peso e no volume em função do comprimento longitudinal. Para essa análise, serão avaliados modelos clássicos amplamente utilizados na modelagem de crescimento biológico, como o Logístico, o modelo de *Von Bertalanffy*, e o modelo de *Richards* explorando suas características para descrever padrões globais e específicos de crescimento. Paralelamente, a análise incluirá modelos de Aprendizado de Máquina, como Árvore de Decisão, *Random Forest* e Máquina de Vetores de Suporte (SVM), visando capturar interações complexas e padrões não lineares que escapam aos modelos clássicos.

A avaliação comparativa entre esses modelos será conduzida com base em critérios estatísticos rigorosos, como o Critério de Informação de Akaike (AIC),o Bayesiano (BIC) e a Raiz do Erro Quadrático Médio (RMSE), possibilitando identificar aqueles que melhor se ajustam aos dados experimentais e oferecem previsões mais precisas. Essa análise permitirá não apenas compreender as dinâmicas gerais do crescimento dos frutos, mas também identificar variações locais significativas que podem impactar a qualidade e produtividade agrícola. Adicionalmente, busca-se explorar as vantagens e limitações de cada abordagem, avaliando, por exemplo, a capacidade do modelo de *Von Bertalanffy* em capturar padrões globais de crescimento e o desempenho do *Random Forest* na previsão de variações locais.

Além disso, a utilização do *software* (Team, 2024) para as análises permitirá a aplicação

de métodos robustos para ajuste de modelos não lineares. A análise dos resíduos será fundamental para garantir a adequação dos modelos e a validade das suposições feitas durante o ajuste, incluindo diagnósticos estatísticos como verificação da normalidade e análise da heterocedasticidade. Isso assegurará que tanto os modelos clássicos quanto os modelos de Aprendizado de Máquina sejam interpretados corretamente, proporcionando previsões confiáveis.

Assim, espera-se que este estudo contribua significativamente para a prática agrícola ao identificar o modelo mais adequado para descrever o crescimento dos frutos de goiabeira "Pedro Sato". A análise comparativa permitirá determinar se os modelos clássicos de regressão não linear, como o de *Von Bertalanffy*, ou os algoritmos de Aprendizado de Máquina, como o *Random Forest*, oferecem melhores resultados em termos de ajuste, previsão e interpretação. Além disso, recomenda-se considerar o uso de tecnologias não destrutivas, como sensoriamento remoto, e incluir variáveis ambientais, como temperatura, umidade e condições do solo, para enriquecer a modelagem e aumentar sua aplicabilidade prática. Esses avanços podem subsidiar decisões relacionadas ao ponto ideal de colheita e à otimização de recursos, promovendo maior eficiência na produção de frutos e melhorias na qualidade final.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção tem como objetivo apresentar os aspectos históricos e conceituais relacionados aos modelos de Regressão não linear, com destaque para os modelos de regressão não linear multiresposta. A discussão é baseada em uma extensa revisão de literatura científica, incluindo livros clássicos, artigos acadêmicos e contribuições fundamentais de pesquisadores da Estatística.

2.1 Marco Histórico

A história da Regressão linear desenvolveu-se no século XIX, quando o matemático e astrônomo Carl Friedrich, como cita Gauss (1809), apresentou o método dos mínimos quadrados. Essa técnica surgiu da necessidade de ajustar dados empíricos a uma reta, permitindo a previsão de valores futuros com base em variáveis observadas. Gauss utilizou o método inicialmente em contextos astronômicos, em que era preciso estimar as órbitas de corpos celestes a partir de observações imprecisas (Stigler, 1986). Simultaneamente, Adrien-Marie, conforme mostra Legendre (1805), também contribuiu para o desenvolvimento desse método, publicando-o de forma independente. O método dos mínimos quadrados minimiza a soma dos quadrados das diferenças entre os valores observados e os valores estimados pela reta de regressão, dada pela equação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,\ \ k = 1, 2, \cdots, k. \tag{2.1}$$

No modelo de regressão linear múltipla, Y representa a variável dependente, $X_1, X_2, \ldots, X_k$ são as variáveis independentes, β_0 é o intercepto (termo constante), $\beta_1, \beta_2, \ldots, \beta_k$ são os coeficientes que representam a contribuição de cada variável independente X_i, e ε é o termo de erro aleatório associado ao modelo. O objetivo do método dos mínimos quadrados é determinar os valores de $\beta_0, \beta_1, \ldots, \beta_k$ que minimizam a soma dos quadrados dos resíduos, ou seja, a soma das diferenças quadradas entre os valores observados y e os valores ajustados $\hat{y}$:

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2. \tag{2.2}$$

Esse método é amplamente utilizado na modelagem linear, pois permite ajustar modelos de forma eficiente, mesmo em casos com várias variáveis explicativas, proporcionando uma aproximação simples e eficaz dos dados observados (Draper; Smith, 1998).

Com o avanço dos estudos estatísticos, ficou claro que a Regressão linear, embora poderosa, possui limitações severas quando os dados não seguem uma relação linear simples. A partir do século XX, pesquisadores começaram a explorar modelos não lineares, uma abordagem que permite modelar relacionamentos mais complexos entre variáveis. O método de Regressão não linear foi formalizado por Ronald A. Fisher, um dos maiores

estatísticos do século XX, que deu um passo além na modelagem estatística e desenvolveu as bases para a análise estatística moderna, inclusive aplicando a metodologia dos mínimos quadrados em contextos não lineares (Fisher, 1925). Além disso, a abordagem dos Mínimos Quadrados Generalizados ou *Generalized Least Squares* (GLS) que é uma extensão do método de Mínimos Quadrados Ordinários ou *Ordinary Least Squares*(OLS), utilizado quando os erros do modelo apresentam heterocedasticidade (variância não constante) ou autocorrelação. Diferente do OLS, que assume que os erros têm variância constante e são independentes, o GLS ajusta as estimativas levando em conta a estrutura de variância e covariância dos erros, proposta por Alexander, como mostra Aitken (1936), surgiu como uma extensão da metodologia de Mínimos Quadrados, sendo utilizada em situações em que as variâncias dos erros não são constantes ou quando há correlação entre as observações. Esses métodos se tornaram ferramentas essenciais em cenários de variáveis interdependentes, como ocorre nos modelos de Regressão não linear multiresposta.

Uma importante extensão dos modelos de Regressão não linear foi o desenvolvimento do modelo de Regressão não linear multiresposta, uma técnica avançada aplicada quando há mais de uma variável dependente (ou resposta) sendo modelada simultaneamente. Esse modelo ganhou notoriedade a partir da segunda metade do século XX, à medida que cientistas de diferentes áreas, como a biologia e a agronomia, começaram a enfrentar situações em que múltiplas respostas interdependentes precisavam ser analisadas simultaneamente (Bates; Watts, 1988).

Inicialmente, o modelo de Regressão não linear multiresposta foi utilizado no campo da química, para ajustar equações de reação complexas com múltiplos produtos (Seber; Wild, 1989). Para lidar com a complexidade desses modelos, várias técnicas de otimização, como o algoritmo de *Gauss-Newton* e o método de *Newton-Raphson*, foram adaptadas. Essas técnicas são iterativas e têm o objetivo de encontrar estimativas para os parâmetros não lineares que melhor se ajustam aos dados (Nocedal; Wright, 2006).

O uso do modelo de Regressão não linear multiresposta expandiu-se para áreas como a engenharia, a ecologia e a biomedicina, no qual a modelagem de múltiplos desfechos é crítica. Por exemplo, no contexto do crescimento de culturas agrícolas, como a mamona, ou na análise de dados econômicos, o modelo pode ser utilizado para prever simultaneamente variáveis como altura, peso, e rendimento, com base em fatores ambientais e de manejo (Ratkowsky, 1990). Além disso, o desenvolvimento de abordagens mais sofisticadas, como o Método dos Momentos Generalizados (GMM, do inglês *Generalized Method of Moments*) é uma técnica econométrica avançada usada para estimar parâmetros de modelos de regressão. Ele foi desenvolvido como uma extensão do método dos momentos e é especialmente útil quando há suspeita de problemas de endogeneidade entre as variáveis explicativas do modelo, proposto por Lars Peter, de acordo com Hansen (1982), permitiu estimativas consistentes em situações em que as suposições tradicionais da regressão, como independência e homoscedasticidade dos erros, são violadas. O (GMM) é

amplamente aplicado em econometria e outros campos que envolvem múltiplas equações simultâneas e respostas correlacionadas, como ocorre frequentemente em modelos multi-resposta.

2.2 Modelo de Regressão não linear

A Regressão não linear é uma classe de modelos em que a relação entre a variável resposta y e as variáveis explicativas x é descrita por uma função linear para não linear nos parâmetros a serem estimados. Este tipo de modelo é essencial quando os dados apresentam comportamentos que não podem ser capturados por uma função linear simples. De acordo com Draper e Smith (1998), a Regressão não linear é amplamente utilizada em áreas como física, biologia, química e economia, sendo uma ferramenta valiosa para a modelagem de processos complexos e dinâmicos que apresentam comportamentos não triviais. A forma geral de um modelo de Regressão não linear pode ser expressa da seguinte maneira:

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon, \quad i = 1, \ldots, n. \tag{2.3}$$

em que, y é a variável dependente, $\mathbf{x}$ é o vetor das variáveis independentes, $\boldsymbol{\theta}$ é o vetor dos parâmetros a serem estimados, $f(\mathbf{x}, \boldsymbol{\theta})$ é a função não linear nos parâmetros $\boldsymbol{\theta}$, e ε é o erro aleatório, assumido normalmente distribuído com $\varepsilon \sim N(0, \sigma^2)$.

A natureza não linear da função $f(\mathbf{x}, \boldsymbol{\theta})$ pode variar de acordo com o fenômeno em estudo, e a escolha dessa função é crucial para garantir a adequação do modelo aos dados. A não linearidade pode estar presente tanto nas variáveis independentes quanto nos parâmetros a serem ajustados.

2.2.1 Modelos de Crescimento

Vários modelos de Regressão não linear (Hastie; Robert; Jerome, 2009) são amplamente utilizados para descrever diferentes tipos de comportamentos em fenômenos reais, como o modelo Logístico, o modelo de *Von Bertalanffy* e o modelo de *Richards*, sendo o modelo Logístico frequentemente utilizado para descrever processos de crescimento limitado, como o crescimento populacional em ecologia que existe uma capacidade de suporte que limita o crescimento de uma população.

Nesse sentido, os modelos de crescimento são ferramentas analíticas fundamentais para descrever e prever como organismos e populações se desenvolvem ao longo do tempo, capturando o comportamento dinâmico e as fases distintas de crescimento. Como observado por Ratkowsky (1990), esses modelos são especialmente úteis para representar processos biológicos e agrícolas, incluindo o desenvolvimento e amadurecimento de frutos, o crescimento de plantas e a expansão de populações animais. De acordo com Von Bertalanffy (1957), o uso de tais modelos permite entender os padrões de crescimento contínuo, adaptando-se bem a organismos cuja taxa de crescimento diminui com o aumento do

tamanho. Além disso, modelos mais flexíveis, como o modelo de *Richards*, introduzem parâmetros adicionais que permitem ajustar a curva de crescimento às particularidades biológicas de cada espécie (Richards, 1959).

Esses modelos não apenas oferecem uma base sólida para quantificar processos de desenvolvimento, mas também possibilitam uma análise detalhada das interações entre ambiente, recursos e características intrínsecas de cada organismo. Como tal, eles têm um valor inestimável tanto para a pesquisa científica quanto para a otimização de práticas agrícolas, ao permitir que se compreenda o impacto de fatores limitantes, como a disponibilidade de nutrientes e espaço, sobre o crescimento (Santos; Silva, 2013).

Modelo Logístico

O modelo Logístico, segundo (Richards, 1959), é um modelo sigmoidal clássico, amplamente utilizado para descrever o crescimento de populações e o desenvolvimento de organismos. Ele captura três fases: uma fase inicial de crescimento lento, uma fase de crescimento acelerado e, por fim, uma fase de estabilização à medida que se aproxima de um valor assintótico máximo. Assim, o modelo pode ser expresso pela fórmula:

$$f(x) = \frac{\alpha}{1 + e^{(\beta - \gamma x)}} \quad \text{para } x \in (-\infty, \infty), \tag{2.4}$$

em que:

- $f(x)$: valor previsto pelo modelo para a variável dependente em função de x;
- α: valor assintótico máximo (capacidade de suporte);
- β: parâmetro relacionado ao ponto de inflexão;
- γ: taxa de crescimento;
- x: variável independente, podendo assumir valores no intervalo real $(-\infty, \infty)$.

O modelo Logístico é especialmente útil para descrever organismos cujo crescimento desacelera com o tempo devido a limitações ambientais, como recursos e espaço.

Modelo de Von Bertalanffy

Proposto por *Ludwig von Bertalanffy* (Von Bertalanffy, 1957), esse modelo é adequado para organismos que continuam a crescer indefinidamente, mas a uma taxa decrescente ao longo do tempo. Frequentemente aplicado a organismos animais e vegetais, ele descreve o crescimento como um processo em que a taxa diminui com o aumento do tamanho do organismo. O modelo é dado pela seguinte fórmula:

$$f(x) = \alpha \cdot \left(1 - e^{-\gamma(x-\delta)}\right) \quad \text{para } x \in (\delta, \infty), \tag{2.5}$$

em que:

- $f(x)$: valor previsto pelo modelo para a variável dependente em função de x;
- α: valor assintótico máximo (tamanho teórico que o organismo pode atingir);
- γ: taxa de crescimento;
- δ: parâmetro relacionado ao tempo inicial (valor mínimo de x para o qual o modelo é válido);
- x: variável independente, válida no intervalo (δ, ∞).

O modelo de *Von Bertalanffy* é especialmente útil para espécies vegetais e animais que exibem crescimento contínuo, ajustando-se bem a curvas de crescimento assimétricas e mais longas.

Modelo de Richards

O modelo de *Richards* (Richards, 1959) é uma extensão do modelo Logístico, permitindo maior flexibilidade através de um parâmetro adicional que controla a forma da curva. Essa flexibilidade o torna aplicável a diversas espécies e contextos biológicos, ajustando tanto crescimentos simétricos quanto assimétricos. O modelo é dado pela seguinte fórmula:

$$f(x) = \frac{\alpha}{1 + e^{(\beta - \gamma x)^{1/\delta}}} \quad \text{para } x \in (\delta, \infty), \tag{2.6}$$

em que:

- $f(x)$: valor previsto pelo modelo para a variável dependente em função de x;
- α: valor assintótico máximo;
- β: parâmetro relacionado ao ponto de inflexão;
- γ: taxa de crescimento;
- δ: parâmetro de forma, que ajusta a simetria da curva (quando $\delta = 1$, o modelo de *Richards* se reduz ao modelo Logístico);
- x: variável independente, válida no intervalo (δ, ∞).

O modelo de *Richards* é especialmente útil para descrever organismos com variações no padrão de crescimento que não são capturadas por um modelo Logístico simples.

Modelo de Gompertz

O modelo de *Gompertz* (Gompertz, 1825) é um modelo sigmoidal, semelhante ao modelo Logístico, mas com uma forma assimétrica. Ele é frequentemente utilizado para descrever o crescimento de tumores, organismos e populações em que o crescimento inicial é rápido, mas a taxa de crescimento diminui de forma exponencial. O modelo é dado pela seguinte fórmula:

$$f(x) = \alpha e^{-e^{(\beta - \gamma x)}} \quad \text{para } x \in (-\infty, \infty), \tag{2.7}$$

em que:

- $f(x)$: valor previsto pelo modelo para a variável dependente em função de x;
- α: valor assintótico máximo (limite superior do crescimento);
- β: parâmetro relacionado ao ponto de inflexão;
- γ: taxa de crescimento;
- x: variável independente, válida no intervalo $(-\infty, \infty)$.

Esse modelo descreve bem processos em que o crescimento se torna muito lento ao se aproximar do valor máximo, como o crescimento de certas frutas e organismos.

Modelo Weibull

O modelo *Weibull* (Weibull, 1951) é amplamente utilizado para representar processos de crescimento e também Análises de Sobrevivência. Ele é flexível e pode ser ajustado para curvas simétricas ou assimétricas, sendo uma alternativa ao modelo de *Richards* quando se precisa de um controle adicional sobre o formato da curva. O modelo é dado pela seguinte fórmula:

$$f(x) = \alpha - \beta e^{-\gamma x^{\delta}} \quad \text{para } x \in (0, \infty), \tag{2.8}$$

em que:

- $f(x)$: valor previsto pelo modelo para a variável dependente em função de x;
- α: valor assintótico máximo (limite superior do crescimento);
- β: parâmetro de escala, que controla o deslocamento da curva;
- γ: taxa de crescimento;
- δ: parâmetro de forma, que ajusta a curvatura da curva (quando $\delta = 1$, o modelo de *Weibull* se reduz ao modelo Exponencial);
- x: variável independente, válida no intervalo $(0, \infty)$.

Esse modelo é útil para descrever o crescimento de organismos e processos em que o aumento é inicialmente lento, acelera e, finalmente, desacelera à medida que se aproxima de um limite assintótico.

Modelo de Morgan-Mercer-Flodin (MMF)

O modelo MMF (Morgan; Mercer; Flodin, 1975) é uma função de crescimento não linear que oferece uma forma flexível para descrever fenômenos de crescimento. Ele é particularmente útil em estudos biológicos e agrícolas devido à sua capacidade de capturar uma variedade de curvas de crescimento, desde simétricas até assimétricas. O modelo é dado pela seguinte fórmula:

$$f(x) = \frac{\beta\gamma + \alpha x^{\delta}}{\gamma + x^{\delta}} \quad \text{para } x \in (0, \infty), \tag{2.9}$$

em que:

- $f(x)$: valor previsto pelo modelo para a variável dependente em função de x;
- α: parâmetro que representa o valor máximo assintótico (limite superior do crescimento);
- β: parâmetro relacionado ao ponto de inflexão da curva;
- γ: taxa de crescimento, controlando o ajuste da curva;
- δ: parâmetro de forma que controla a curvatura da curva;
- x: variável independente, válida no intervalo $(0, \infty)$.

O modelo MMF é vantajoso em estudos de crescimento onde é importante capturar tanto a fase de crescimento inicial quanto a estabilização de forma precisa, oferecendo um ajuste para situações em que a curva apresenta uma inflexão mais controlada em comparação ao modelo Logístico.

Os modelos de crescimento não linear, como o Logístico, *Von Bertalanffy*, *Richards*, *Gompertz*, *Weibull* e *Morgan, Mercer e Flodin* (MMF), têm sido amplamente utilizados em diversas áreas da biologia, agronomia e ecologia, como demonstrado por estudos como o de Zeviani, Ribeiro Júnior e Bonat (2013), que aplicaram o modelo de *Von Bertalanffy* no estudo de crescimento de frutas, e o trabalho de Hastie, Robert e Jerome (2009), que abordou os diferentes padrões de crescimento utilizando modelos logísticos.

Contudo, a linearização desses modelos é frequentemente necessária para a estimação de parâmetros, como observam Draper e Smith (1998) em sua análise sobre regressão não linear. Embora a linearização, por meio de transformações logarítmicas ou recíprocas, permita o uso de métodos de Mínimos Quadrados, ela pode distorcer as dinâmicas não

lineares reais dos dados, resultando em ajustes menos precisos (Seber; Lee, 2003). Isso se aplica especialmente quando as características do crescimento são complexas e envolvem interações não capturadas pelas transformações.

Em alguns casos, a linearização pode não ser adequada para lidar com a variabilidade dos dados ou com violações dos pressupostos, como a homocedasticidade, o que justifica a necessidade de métodos mais avançados, como os de Mínimos Quadrados Generalizados (GLS) ou técnicas de aprendizado de máquina, como discutido por Hastie, Robert e Jerome (2009). Esses métodos podem proporcionar melhores ajustes em situações onde os modelos lineares não são capazes de capturar toda a complexidade do crescimento biológico.

2.3 Modelos de Regressão Multiresposta

Os modelos de Regressão multiresposta (Fogliatto, 2008) ampliam o conceito de regressão para situações em que há múltiplas variáveis dependentes que precisam ser ajustadas simultaneamente em relação a um conjunto comum de preditores. Esses modelos são amplamente utilizados em áreas como biologia, química e economia, em que diferentes variáveis de resposta estão correlacionadas entre si e dependem das mesmas variáveis explicativas.

Seja $\mathbf{Y}$ um vetor de m variáveis dependentes (respostas) e X uma matriz de p variáveis independentes (preditores). O modelo de regressão multiresposta é dado por:

$$\mathbf{Y} = f\left(\mathbf{X},\ \boldsymbol{\Theta}\right) + \boldsymbol{\varepsilon}, \tag{2.10}$$

em que, $\mathbf{Y}$ é o vetor de respostas $\mathbf{Y} = (y_1, y_2, \ldots, y_m)$, $\mathbf{X}$ é a matriz dos preditores $\mathbf{X} = (x_1, x_2, \ldots, x_p)$, Θ é a matriz de parâmetros $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots, \theta_m)$, $\boldsymbol{\varepsilon}$ é o vetor de erros, assumido normalmente distribuído com $\boldsymbol{\varepsilon} \sim N(0, \Sigma)$, no qual Σ é a matriz de covariância entre os resíduos.

A estimação dos parâmetros em modelos multiresposta pode ser realizada utilizando métodos iterativos, como o método de *Gauss-Newton* multivariado, que generaliza o método de *Gauss-Newton* para múltiplas respostas. A equação de atualização dos parâmetros é dada por:

$$\boldsymbol{\Theta}_{k+1} = \boldsymbol{\Theta}_k - \left(\mathbf{J}^T\mathbf{J}\right)^{-1}\mathbf{J}^T\mathbf{r}, \tag{2.11}$$

em que, $\mathbf{J}$ é a matriz Jacobiana, que contém as derivadas parciais das funções de resposta com relação aos parâmetros, e $\mathbf{r}$ é o vetor de resíduos multivariados. Como afirmado por Bates e Watts (1988), esse método é eficiente quando as funções de resposta são suavemente não lineares e os parâmetros podem ser aproximados de maneira iterativa.

A utilização de modelos multiresposta permite capturar a interdependência entre as diferentes variáveis de resposta, proporcionando ajustes mais precisos e interpretativos. No entanto, a complexidade do ajuste aumenta significativamente, especialmente devido

à necessidade de estimar a matriz de covariância S, que representa a correlação entre os resíduos das diferentes respostas.

A avaliação da qualidade do ajuste nesses modelos deve ser realizada de forma abrangente, utilizando tanto métricas numéricas quanto análise gráfica dos resíduos. Medidas como o Coeficiente de Determinação (R^2) (Hastings, 2000), o Erro Quadrático Médio da Raiz (RMSE) (Willmott; Matsuura, 2005) e o Critério de Informação de Akaike (AIC) (Akaike, 1974) são úteis para quantificar o ajuste e comparar diferentes modelos, enquanto a análise de resíduos fornece informações importantes sobre possíveis violações de suposições.

Para situações envolvendo múltiplas variáveis dependentes, os modelos de Regressão multiresposta oferecem uma solução robusta, mas que exige cuidados na estimação dos parâmetros e na interpretação dos resultados. O sucesso desses modelos depende da escolha adequada das variáveis, do método de estimação e da correta modelagem das correlações entre os resíduos.

2.4 Métodos de Estimação

A estimação dos parâmetros θ em modelos de regressão não linear é geralmente feita através de métodos iterativos, uma vez que não é possível obter soluções analíticas exatas como na regressão linear. Entre os métodos mais comuns para esse ajuste, destacam-se o Método dos Mínimos Quadrados não Lineares e o Método da Máxima Verossimilhança.

2.4.1 Mínimos Quadrados Não Lineares

O método de estimação por Mínimos Quadrados não Lineares (Gauss, 1809) é uma técnica amplamente empregada para ajustar modelos não lineares, estimando os parâmetros θ que minimizam a Soma dos Quadrados dos Resíduos. Esse método é fundamental quando a relação entre as variáveis explicativas e a variável resposta não pode ser descrita de maneira linear, tornando-se necessário recorrer a funções não lineares para capturar a complexidade dos dados.

O resíduo r_i é a diferença entre o valor observado da variável resposta y_i e o valor predito pelo modelo $f(x_i, \theta)$. Matematicamente, pode-se expressar o resíduo como:

$$r_i = y_i - f(x_i, \theta), \tag{2.12}$$

em que, y_i é o valor observado da variável resposta; $f(x_i, \theta)$ é o valor ajustado pelo modelo para o conjunto de parâmetros θ e variáveis explicativas x_i.

O objetivo do método é encontrar o vetor de parâmetros θ que minimize a Soma dos Quadrados dos Resíduos, conforme descrito a seguir. A função objetivo dos Mínimos Quadrados não Lineares é a minimização da Soma dos Quadrados dos Resíduos $S(\theta)$,

dada por:

$$S(\theta) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} [y_i - f(x_i, \theta)]^2. \tag{2.13}$$

Neste caso, n é o número total de observações; r_i representa o resíduo associado à i-ésima observação; A função $f(x_i, \theta)$ descreve a forma funcional que representa a relação entre as variáveis independentes x e a variável dependente y, parametrizada por θ.

Esse processo de minimização busca ajustar o modelo de forma a reduzir a discrepância entre os valores observados y_i e os valores ajustados pelo modelo $f(x_i, \theta)$, resultando no melhor ajuste possível aos dados. Assim, o vetor de resíduos r pode ser expresso de maneira compacta como:

$$r = \mathbf{y} - f(x, \theta), \tag{2.14}$$

em que, $\mathbf{y}$ é o vetor de todos os valores observados; $f(x, \theta)$ é o vetor correspondente aos valores ajustados pelo modelo. Esse vetor é central no processo de otimização, pois captura a diferença entre o modelo e os dados reais. O objetivo do método de Mínimos Quadrados é ajustar θ de forma que os elementos desse vetor sejam minimizados.

Para minimizar a Soma dos Quadrados dos Resíduos $S(\theta)$, a função é derivada em relação aos parâmetros θ. O sistema de equações resultante pode ser expresso como:

$$\frac{\partial S(\theta)}{\partial \theta} = -2 \sum_{i=1}^{n} [y_i - f(x_i, \theta)] \frac{\partial f(x_i, \theta)}{\partial \theta} = 0. \tag{2.15}$$

Esse sistema representa um conjunto de equações não lineares em θ, as quais precisam ser resolvidas iterativamente, uma vez que soluções analíticas não estão disponíveis para a maioria dos modelos não lineares. O método baseia-se em encontrar um conjunto de parâmetros θ que satisfaça essas equações, resultando no melhor ajuste possível.

A resolução desse sistema de equações não lineares geralmente requer o uso de métodos numéricos iterativos. Dois dos métodos mais comuns são o Método de *Gauss-Newton* e o Método de *Newton-Raphson* (Nocedal; Wright, 2006). Ambos os métodos iteram sucessivamente sobre os valores de θ até que a função objetivo $S(\theta)$ atinja um mínimo, dentro de uma tolerância aceitável.

A convergência desses métodos depende de várias condições, incluindo a escolha de um bom ponto inicial para θ e a suavidade da função $f(x, \theta)$. Em muitos casos, escolhas inadequadas podem levar a uma convergência lenta ou até à divergência do algoritmo. Por isso, é comum utilizar técnicas como regularização ou *damping* para garantir que o processo de otimização seja estável.

2.4.2 Método da Máxima Verossimilhança

O método de estimação por Máxima Verossimilhança (EMV), introduzido por Ronald, como aponta Fisher (1922), tornou-se uma das abordagens mais fundamentais e ampla-

mente utilizadas para a estimativa de parâmetros em modelos estatísticos. O objetivo central desse método é encontrar o conjunto de parâmetros θ que maximize a função de verossimilhança, a qual representa a probabilidade de observar um determinado conjunto de dados, dado um modelo específico. De acordo com Casella e Berger (2002), o método de Máxima Verossimilhança é particularmente atraente devido à sua assintótica eficiência, o que implica que, com um número suficientemente grande de observações, as estimativas de máxima verossimilhança são as mais precisas e consistentes possíveis.

A função de verossimilhança para um conjunto de dados $y_1, y_2, \ldots, y_n$, dado um modelo parametrizado por θ, é definida como o produto das funções de densidade de probabilidade (ou funções de massa de probabilidade, no caso de variáveis discretas) dos dados observados.

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta), \tag{2.16}$$

em que, $f(y_i; \theta)$ é a função de densidade de probabilidade do i-ésimo dado observado y_i, condicionada aos parâmetros θ; $L(\theta)$ é a verossimilhança conjunta dos dados, dada a suposição de que as observações são independentes.

A função de verossimilhança expressa a probabilidade de observar o conjunto de dados $y_1, y_2, \ldots, y_n$, assumindo que o modelo com parâmetros θ seja o correto. O objetivo do Método de Máxima Verossimilhança é encontrar o valor de θ que maximiza essa função.

Na prática, para simplificar o processo de maximização, é comum utilizar a função log-verossimilhança. Isso se deve ao fato de que o logaritmo transforma o produto das densidades de probabilidade em uma soma, facilitando tanto a interpretação quanto os cálculos. A função log-verossimilhança é dada por:

$$l(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta). \tag{2.17}$$

Maximizar a log-verossimilhança é equivalente a maximizar a função de verossimilhança, já que o logaritmo natural é uma função monotonicamente crescente. Portanto, o valor de θ que maximiza $l(\theta)$ também maximiza $L(\theta)$.

O valor de θ que maximiza a função log-verossimilhança $l(\theta)$ é chamado de estimativa de Máxima Verossimilhança (EMV). Matematicamente, isso é expresso como a solução para o sistema de equações dado pela derivada da função log-verossimilhança em relação aos parâmetros θ, que deve ser igual a zero:

$$\frac{\partial l(\theta)}{\partial \theta} = 0. \tag{2.18}$$

Esse sistema de equações é geralmente não linear e, portanto, exige o uso de métodos numéricos iterativos para encontrar a solução. Entre os métodos mais comuns estão o método de *Newton-Raphson* e o método de Gradiente ascendente. Ambos utilizam

aproximações sucessivas dos parâmetros θ até que se atinja a convergência, ou seja, até que a função log-verossimilhança seja maximizada de forma satisfatória.

Uma das propriedades mais importantes das estimativas de máxima verossimilhança é sua eficiência assintótica, conforme destacado por (Casella; Berger, 2002). Isso significa que, à medida que o número de observações n aumenta, as estimativas de θ obtidas pelo método de Máxima Verossimilhança tornam-se consistentes e atingem a menor variância possível entre todas as estimativas não tendenciosas. Em outras palavras, com um grande número de dados, o método de Máxima Verossimilhança oferece estimativas que são, em média, as melhores possíveis.

Para obter as estimativas de Máxima Verossimilhança de θ_1, θ_2 e σ^2, é necessário maximizar a função de verossimilhança em relação a esses parâmetros. No entanto, maximizar diretamente a função de verossimilhança pode ser complexo devido à sua forma não linear. Por isso, é comum utilizar a função log-verossimilhança, que neste caso se torna:

$$l\left(\theta_1, \theta_2, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \theta_1 e^{-\theta_2 x_i}\right)^2. \tag{2.19}$$

Maximizar essa função envolve a resolução numérica de um sistema de equações não lineares. Ferramentas computacionais, como o algoritmo de *Newton-Raphson* ou métodos baseados em gradiente, são frequentemente utilizadas para encontrar as soluções ótimas para θ_1, θ_2 e σ^2.

2.5 Métodos de Otimização

Os métodos de otimização são cruciais no ajuste de modelos de Regressão não lineares, visto que a estimativa dos parâmetros requer a minimização de uma função objetivo, como a Soma dos Quadrados dos Resíduos ou a função de verossimilhança (Draper; Smith, 1998). A natureza não linear de muitos problemas de regressão exige técnicas iterativas para encontrar soluções que melhor se ajustem aos dados observados. A seguir, serão discutidos alguns dos métodos de otimização mais utilizados em tais contextos, com destaque para os métodos de *Gauss-Newton*, Quase-*Newton* e *Newton-Raphson*.

2.5.1 Método de *Gauss-Newton*

O método de *Gauss-Newton* (Gauss, 1809) é um dos algoritmos mais comuns para resolver problemas de mínimos quadrados não lineares. Ele parte da premissa de que a função $f(x, \theta)$ pode ser linearizada em torno de uma aproximação inicial θ_0, utilizando uma expansão de primeira ordem da série de *Taylor*. Isso leva a um problema de Mínimos Quadrados Linearizado, cuja solução é obtida de forma iterativa. A atualização dos parâmetros no método de *Gauss-Newton* é dada por:

$$\theta_{k+1} = \theta_k - \left(J^T J\right)^{-1} J^T \mathbf{r}, \tag{2.20}$$

em que ,$f_t(x,\theta)$ com respeito aos parâmetros θ; $\mathbf{r}$ é o vetor de resíduos $\mathbf{r} = y - f(x,\theta)$, em que y são os valores observados e $f(x,\theta)$ são os valores ajustados pelo modelo.

Este método, conforme discutido por Bates e Watts (1988), é bastante eficiente quando a função $f(x,\theta)$ é aproximadamente linear nos parâmetros. Em tais casos, o método converge rapidamente. No entanto, quando a função é fortemente não linear, o método pode falhar ou apresentar convergência lenta, exigindo uma boa escolha inicial dos parâmetros para funcionar adequadamente.

2.5.2 Método Quase-Newton

Os métodos Quase-*Newton* (Davidon, 1959) formam uma classe de algoritmos de otimização que utilizam aproximações da matriz Hessiana (a matriz de segundas derivadas) em vez de calcular a Hessiana exata. O algoritmo BFGS (*Broyden-Fletcher-Goldfarb-Shanno*) (Broyden et al., 1970) é o mais popular desta classe e atualiza iterativamente a aproximação da Hessiana utilizando informações das derivadas de primeira ordem.

A atualização da matriz Hessiana no método BFGS é dada por:

$$H_{k+1} = H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k^T}{s_k^T H_k s_k}, \tag{2.21}$$

em que, $s_k = \theta_{k+1} - \theta_k$ é a diferença entre as estimativas dos parâmetros em duas iterações sucessivas; $y_k = \nabla f(\theta_{k+1}) - \nabla f(\theta_k)$ é a diferença entre os gradientes da função objetivo em duas iterações consecutivas.

Conforme apontado por Nocedal e Wright (2006), o método BFGS é mais robusto que o método de *Gauss-Newton*, pois não depende da linearização da função objetivo. Além disso, ele pode ser aplicado com sucesso em uma ampla gama de problemas, incluindo aqueles em que a função objetivo é fortemente não linear ou mal condicionada. A vantagem principal desse método é que ele combina a precisão dos métodos baseados em segundas derivadas com a eficiência computacional dos métodos baseados em gradiente.

2.5.3 Método de *Newton-Raphson*

O método de*Newton-Raphson* (Burden; Faires, 2016) é um método iterativo de otimização que utiliza a matriz Hessiana para encontrar o ponto de mínimo ou máximo de uma função objetivo. O princípio básico do método é que, ao utilizar uma aproximação de segunda ordem da função objetivo, pode-se melhorar significativamente a taxa de convergência. A atualização dos parâmetros é dada pela equação:

$$\theta_{k+1} = \theta_k - H^{-1} \nabla S(\theta_k), \tag{2.22}$$

em que, H é a matriz Hessiana, ou matriz de segundas derivadas, da função objetivo $S(\theta)$; $\nabla S(\theta_k)$ é o gradiente da função objetivo no ponto θ_k.

De acordo com Seber e Wild (1989), o método de *Newton-Raphson* é extremamente eficiente, desde que a função objetivo seja suave e a aproximação inicial dos parâmetros esteja relativamente próxima do ponto ótimo. Em tais situações, o método pode convergir muito rapidamente para a solução ideal. No entanto, quando a aproximação inicial está longe da solução, ou se a função não é suficientemente suave, o método pode divergir, levando a soluções errôneas ou exigindo a introdução de técnicas de regularização.

Cada um desses métodos de otimização tem suas próprias vantagens e desvantagens. O método de *Gauss-Newton* é ideal para funções quase lineares, mas sua aplicação é limitada quando a função é fortemente não linear. O método Quase-*Newton*, por outro lado, é mais robusto, pois utiliza aproximações da matriz Hessiana sem a necessidade de calculá-la explicitamente, o que o torna eficiente em uma ampla gama de problemas. Já o método de *Newton-Raphson* oferece convergência rápida em problemas bem comportados, mas é sensível a aproximações iniciais e à suavidade da função objetivo.

2.6 Qualidade do Ajuste

Avaliar a qualidade do ajuste em modelos de Regressão não linear é um passo essencial para garantir que o modelo escolhido captura corretamente as relações entre as variáveis. De forma geral, essa avaliação envolve uma análise cuidadosa dos parâmetros estimados, bem como da capacidade do modelo em explicar a variabilidade observada nos dados. É fundamental utilizar uma combinação de critérios quantitativos e métodos gráficos para verificar a adequação do modelo. Alguns dos principais métodos incluem o Coeficiente de Determinação (R^2) (Hastings, 2000), a Análise de Resíduos, e Testes de Hipóteses para os parâmetros do modelo. Além disso, medidas de ajuste como o Erro Quadrático Médio da Raíz(RMSE) (Willmott; Matsuura, 2005) e o Critério de Informação de Akaike (AIC) (Akaike, 1974) podem ser úteis para comparar diferentes modelos.

2.6.1 Coeficiente de Determinação

O Coeficiente de Determinação R^2 é um dos indicadores mais utilizados para medir a qualidade do ajuste de modelos de regressão (Hastings, 2000). Ele indica a proporção da variabilidade total na variável dependente Y que é explicada pelo modelo ajustado. A fórmula do R^2 é dada por:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}, \tag{2.23}$$

em que y_i são os valores observados, $\widehat{y_i}$ são os valores ajustados pelo modelo, e $\bar{y}$ é a média dos valores observados.

De acordo com Draper e Smith (1998), um valor de R^2 próximo de 1 indica que o modelo explica a maior parte da variabilidade dos dados, sugerindo um bom ajuste.

No entanto, é importante salientar que, em modelos não lineares, o R^2 pode não ser tão interpretável quanto nos modelos lineares, especialmente quando a relação entre as variáveis não segue um comportamento simples. Além disso, o R^2 pode ser inflacionado quando se adicionam mais preditores ao modelo, mesmo que eles não sejam relevantes. Por isso, medidas ajustadas, como o R^2 ajustado, podem ser mais apropriadas para evitar essa distorção. Assim,

$$R^2_{\text{ajustado}} = 1 - \left(\frac{n-1}{n-p-1}\right) \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{2.24}$$

Aqui, n é o número de observações e p é o número de parâmetros estimados no modelo. O R^2 ajustado penaliza o uso de variáveis irrelevantes, sendo uma métrica mais confiável para comparar modelos de diferentes complexidades (Seber; Lee, 2003).

No entanto, para modelos não lineares, como os modelos de regressão logística ou outros modelos de aprendizado de máquina, o R^2 tradicional pode não ser adequado. Nesses casos, o pseudo-R^2 é uma alternativa. Existem diferentes versões de pseudo-R^2, uma das mais conhecidas é a de *McFadden*, que é dada por:

$$R^2_{\text{McFadden}} = 1 - \frac{\ln(L_{\text{modelo}})}{\ln(L_{\text{nulo}})}, \tag{2.25}$$

em que L_{modelo} é a verossimilhança do modelo ajustado e L_{nulo} é a verossimilhança do modelo nulo, ou seja, o modelo sem variáveis independentes. O pseudo-R^2 de McFadden é uma medida popular para modelos logísticos e indica a melhoria da verossimilhança em relação ao modelo nulo. Quanto mais próximo de 1 for o valor de R^2_{McFadden}, melhor será o ajuste do modelo aos dados. É uma alternativa útil quando a relação entre as variáveis não é linear.

Esses diferentes coeficientes de determinação permitem avaliar a qualidade do ajuste de diferentes tipos de modelos e fornecer uma compreensão mais robusta sobre o desempenho dos modelos ajustados, especialmente quando lidamos com modelos não lineares ou de aprendizado de máquina.

2.6.2 Análise de Resíduos

A Análise de Resíduos é um componente fundamental na avaliação da adequação do modelo. Os resíduos r_i são definidos como a diferença entre os valores observados y_i e os valores ajustados $\widehat{y_i}$ pelo modelo, isto é:

$$r_i = y_i - \widehat{y_i}. \tag{2.26}$$

De acordo com Seber e Lee (2003), para que o modelo seja considerado adequado, os resíduos devem ser distribuídos aleatoriamente em torno de zero, sem apresentar padrões

ou tendências específicas. Se os resíduos mostram qualquer tipo de padrão sistemático (como heterocedasticidade ou autocorrelação), isso pode indicar que o modelo não está capturando corretamente a estrutura dos dados. A inspeção gráfica dos resíduos, como o gráfico de resíduos *versus* valores ajustados, é uma prática comum para identificar problemas de ajuste. Se os resíduos apresentam uma dispersão crescente ou decrescente em relação aos valores ajustados, isso sugere heterocedasticidade, ou seja, uma variação não constante dos erros ao longo dos níveis de X.

Testes formais, como o teste de *Shapiro-Wilk* (Shapiro; Wilk, 1965) podem ser utilizados para verificar a normalidade dos resíduos. Caso o teste indique uma distribuição não normal, ajustes no modelo podem ser necessários para melhorar o ajuste. Outro teste formal, o teste de *Breusch-Pagan*, pode ser utilizado para detectar heterocedasticidade nos resíduos. Este teste verifica se a variância dos resíduos depende dos valores ajustados. Um resultado significativo indicaria a presença de heterocedasticidade, sugerindo que o modelo precisa ser ajustado, através da transformação dos dados ou da utilização de modelos que acomodem erros com variância não constante, como os modelos de Regressão ponderada.

Além disso, o teste de *Durbin-Watson* é frequentemente utilizado para detectar autocorrelação nos resíduos, especialmente em dados temporais. A autocorrelação positiva ou negativa dos resíduos sugere que o modelo pode não ter capturado adequadamente a dependência temporal ou a estrutura de correlação presente nos dados (Seber; Lee, 2003).

2.6.3 Critério de Informação de Akaike (AIC) e Bayesiano (BIC) e Raiz do Erro Quadrático Médio (RMSE)

Além das análises de R^2 e resíduos, outras métricas são essenciais para avaliar a qualidade do ajuste e comparar diferentes modelos. O Critério de Informação de Akaike (AIC) é uma medida que considera tanto a qualidade do ajuste quanto a simplicidade do modelo. Dessa forma:

$$AIC = 2p - 2\ln(L), \tag{2.27}$$

no qual p é o número de parâmetros no modelo e L é o valor da função de verossimilhança do modelo ajustado. Como explicado por Akaike (1974), o AIC é uma medida relativa, útil para comparar modelos com diferentes números de parâmetros, penalizando aqueles que são excessivamente complexos em relação à melhoria no ajuste. Essa abordagem é corroborada por Burnham e R. (2002), que ressaltam que o AIC deve ser utilizado em contextos que se busca o modelo que melhor explica os dados.

Da mesma forma, o Critério de Informação Bayesiano (BIC) também é utilizado para essa finalidade e é calculado como:

$$BIC = \ln(n)p - 2\ln(L), \tag{2.28}$$

em que n é o número de observações. O BIC, assim como o AIC, considera a qualidade do ajuste, mas aplica uma penalização maior para modelos com mais parâmetros, favorecendo modelos mais simples Schwarz (1978).

Além disso, o Erro Quadrático Médio da Raiz (RMSE) (Willmott; Matsuura, 2005) é uma métrica relevante para avaliar a precisão do modelo. O RMSE é definido como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}. \tag{2.29}$$

O RMSE fornece uma medida em unidades da variável predita, facilitando a interpretação do erro de previsão, já que valores menores de RMSE indicam melhor ajuste do modelo aos dados observados. Essa métrica é frequentemente utilizada em estudos de modelagem, como descrito por Hyndman e B. (2006), que enfatizam sua eficácia em mensurar a precisão das previsões.

Juntas, essas métricas proporcionam uma visão abrangente da qualidade do ajuste dos modelos, permitindo decisões informadas na seleção do modelo mais adequado para os dados analisados.

2.7 Aprendizado de máquina (*Machine Learning*)

Segundo Alpaydin (2020), o campo do Aprendizado de Máquina (ou *Machine Learning*) refere-se ao desenvolvimento de algoritmos que permitem que computadores aprendam a partir de dados, sem a necessidade de programação explícita. Esse processo é possível por meio de modelos matemáticos que, ao detectar padrões nos dados de entrada, permitem ao sistema realizar previsões ou classificações baseadas em novas informações (Goodfellow, 2016). Desse modo, algoritmos de Aprendizado de Máquina vão além de uma programação rígida para seguir comandos fixos; eles são desenvolvidos para se adaptar e aprimorar continuamente conforme recebem mais dados (Russell; Peter, 2020).

O aprendizado de máquina é comumente dividido em três categorias principais:

i) **Aprendizado Supervisionado**: O modelo é treinado com dados rotulados, onde há uma associação definida entre entradas e saídas. Como aponta Murphy (2012), esse método é amplamente empregado para tarefas não supervisionado e de reforço, pois permite que o modelo aprenda padrões específicos e consiga aplicá-los a dados desconhecidos.

ii) **Aprendizado Não Supervisionado**: Nesse contexto, o modelo trabalha com dados sem rótulos, explorando-os para identificar padrões internos. Segundo Sutton e G. (2018), o aprendizado não supervisionado é essencial em casos onde a relação direta entre entrada e saída não é conhecida.

iii) **Aprendizado por Reforço**: Um agente interage com o ambiente e aprende a maximizar a recompensa acumulada, ajustando suas ações conforme recebe o *feedback*. Como explica Sutton e G. (2018), esse método baseia-se no aprendizado comportamental, onde o agente é incentivado pela maximização de recompensas.

O Aprendizado de Máquina possui aplicações em várias áreas, como diagnóstico médico, reconhecimento de padrões e análise de dados financeiros. Conforme destacado por Goodfellow (2016), essa tecnologia tem o potencial de revolucionar inúmeros setores, oferecendo soluções ágeis e eficientes para problemas complexos e de grande escala.

2.7.1 Árvore de Decisão

O modelo de Árvore de Decisão é um dos métodos mais intuitivos e interpretáveis em Aprendizado de Máquina, frequentemente utilizado tanto em problemas de classificação quanto de regressão. As árvores de decisão funcionam criando uma estrutura hierárquica de regras baseadas nas variáveis de entrada, em que cada nó representa uma decisão ou condição, e as folhas representam os resultados finais.

De acordo com Hastie, Robert e Jerome (2009), o modelo de Árvore de Decisão é desenvolvido por meio de sucessivas divisões nos dados de treinamento, baseadas nos valores das variáveis preditoras, criando um percurso que culmina em uma previsão final. Esse modelo é especialmente útil em problemas em que a interpretabilidade é essencial, já que sua estrutura hierárquica facilita a visualização do processo de tomada de decisão. O funcionamento do modelo é o seguinte:

i) **Divisão Recursiva**: O processo começa com a divisão do conjunto de dados com base na variável que melhor separa as classes (no caso de classificação) ou que minimiza o erro (no caso de regressão), repetindo-se de maneira recursiva até que um critério de parada seja atingido (Breiman; Friedman et al., 1984).

ii) **Critério de Parada**: Uma árvore de decisão pode continuar se ramificando até que todas as folhas contenham exemplos puros ou um número mínimo de amostras. Esse critério evita o sobreajuste e controla o tamanho da árvore (Breiman; Friedman et al., 1986).

iii) **Predição**: A predição para uma nova amostra é feita ao seguir o caminho da árvore, tomando decisões em cada nó até chegar em uma folha com a predição final (Hastie; Robert; Jerome, 2009).

Vantagens da Árvore de Decisão:

i) **Interpretabilidade**: A estrutura da árvore facilita a compreensão do processo decisório, tornando o modelo valioso em áreas onde é fundamental justificar as decisões tomadas (Hastie; Robert; Jerome, 2009).

ii) **Versatilidade**: Pode lidar com dados categóricos e contínuos, e não requer suposições sobre a distribuição dos dados (Breiman; Friedman et al., 1986).

iii) **Rápido Treinamento**: As árvores de decisão são rápidas de treinar, sendo eficientes para dados de tamanho pequeno a médio (Breiman; Friedman et al., 1986).

Desvantagens da Árvore de Decisão:

i) **Tendência ao Sobreajuste**: Árvores muito complexas podem se ajustar demais aos dados de treinamento, resultando em baixa capacidade de generalização. Técnicas como poda ou uso de métodos em conjunto, como *Random Forest*, podem mitigar esse problema (Hastie; Robert; Jerome, 2009).

ii) **Sensibilidade a Variações nos Dados**: Árvores de Decisão podem ser instáveis, alterando-se significativamente com pequenas variações nos dados.

O modelo de Árvore de Decisão é amplamente utilizado em áreas como diagnóstico médico e crédito bancário, em que a interpretabilidade e a explicabilidade das decisões são importantes.

2.7.2 Máquina de Vetores de Suporte (SVM)

O modelo de Máquina de Vetores de Suporte (SVM) é uma técnica poderosa e eficaz em Aprendizado de Máquina, particularmente em problemas de classificação binária, mas também aplicável à regressão. As SVMs funcionam identificando um hiperplano que separa as classes de forma ótima, maximizando a margem entre os dados de classes distintas. Essa margem máxima ajuda a melhorar a generalização do modelo.

De acordo com Cortes e Vapnik (1995), a Máquina de Vetores de Suporte (SVM) busca identificar um hiperplano ótimo que maximize a margem entre as classes, considerando apenas os dados mais próximos à fronteira de decisão, conhecidos como vetores de suporte. Esse processo confere ao modelo um desempenho sólido, especialmente em problemas de alta dimensionalidade e onde a separação entre classes não é linear. Esse método consiste em:

i) **Maximização da Margem**: O objetivo da Máquina de Vetores de Suporte (SVM) é encontrar o hiperplano que maximiza a distância (margem) entre as classes de dados, reduzindo a possibilidade de erro de classificação (Cortes; Vapnik, 1995).

ii) **Vetores de Suporte**: Somente os dados próximos à margem, os vetores de suporte, são relevantes para definir o hiperplano, o que ajuda a reduzir a complexidade do modelo (Schölkopf; Smola, 2002).

iii) **Truque do Kernel**: Para dados que não são linearmente separáveis, a Máquina de Vetores de Suporte (SVM) utiliza funções de *kernel* para projetar os dados em um espaço dimensional maior, onde um hiperplano linear pode separá-los (Vapnik, 1998).

Vantagens da Máquina de Vetores de Suporte (SVM):

i) **Eficiente em Alta Dimensionalidade**: A Máquina de Vetores de Suporte (SVM) lida bem com dados de alta dimensionalidade e é eficaz em problemas com grande número de variáveis (Cortes; Vapnik, 1995).

ii) **Robustez a Outliers**: Como apenas os vetores de suporte afetam o modelo final, a Máquina de Vetores de Suporte (SVM) é menos sensível a *outliers*.

iii) **Flexibilidade com Kernels**: A possibilidade de usar diferentes tipos de *kernels* torna a Máquina de Vetores de Suporte (SVM) altamente versátil e aplicável a uma ampla variedade de problemas não lineares (Schölkopf; Smola, 2002).

Desvantagens da Máquina de Vetores de Suporte (SVM):

i) **Complexidade Computacional**: O treinamento de uma Máquina de Vetores de Suporte (SVM) pode ser computacionalmente caro, especialmente em grandes conjuntos de dados (Cortes; Vapnik, 1995).

ii) **Dificuldade de Interpretação**: Diferente de modelos como árvores de decisão, a Máquina de Vetores de Suporte (SVM) não é facilmente interpretável, o que pode dificultar a compreensão dos critérios de classificação (Hastie; Robert; Jerome, 2009).

As Máquinas de Vetores de Suporte (SVM) são amplamente utilizadas em aplicações como classificação de imagens e reconhecimento de padrões, áreas em que a precisão e a capacidade de lidar com dados complexos são fundamentais.

2.7.3 Random Forest

O modelo *Random Forest* (ou Floresta Aleatória) é um dos métodos mais populares e robustos em Aprendizado de Máquina supervisionado, sendo amplamente utilizado para problemas de classificação e regressão. Ele pertence aos chamados *ensemble methods* (métodos de comitê), que combinam múltiplos modelos para aumentar a precisão e a estabilidade das previsões (Breiman, 2001).

Segundo Breiman (2001), o *Random Forest* consiste na criação de "múltiplas árvores de decisão a partir de subconjuntos aleatórios dos dados de treinamento, em que cada árvore contribui para a predição final, resultando em um modelo mais preciso e menos suscetível

ao sobreajuste" (p. 6). Esse método aproveita a diversidade das árvores individuais e melhora o desempenho geral ao combinar suas previsões, funcionando da seguinte forma:

i) **Construção de Árvores de Decisão**: O *Random Forest* utiliza subconjuntos aleatórios dos dados de treinamento para criar um conjunto de árvores de decisão independentes. "Essa técnica, conhecida como *bagging*, ajuda a reduzir a variância do modelo e melhora a estabilidade das previsões" (Hastie; Robert; Jerome, 2009).

ii) **Seleção Aleatória de Variáveis**: Em cada nó da árvore, um subconjunto aleatório de variáveis é escolhido para dividir os dados. De acordo com Breiman (2001), isso "introduz uma diversidade adicional nas árvores, promovendo uma maior generalização do modelo".

iii) **Predição**: Ao realizar previsões, todas as árvores do modelo são consultadas, e suas respostas são combinadas. Para classificação, o resultado final é obtido por voto majoritário; para regressão, pela média das predições individuais (Hastie; Robert; Jerome, 2009).

Vantagens do Random Forest:

i) **Robustez ao Sobreajuste**: Ao combinar os resultados de várias árvores, o *Random Forest* minimiza a probabilidade de sobreajuste, pois "a média dos resultados das árvores individuais ajuda a reduzir a variância do modelo" (Breiman, 2001).

ii) **Alta Precisão**: É especialmente adequado para dados de alta dimensionalidade e permite identificar variáveis mais importantes, o que "contribui para o aumento da precisão em problemas complexos" (Hastie; Robert; Jerome, 2009).

iii) **Versatilidade**: O modelo é flexível e pode ser utilizado com variáveis categóricas e contínuas, facilitando sua aplicação em uma variedade de problemas.

Desvantagens do Random Forest:

i) **Demanda Computacional**: A construção de muitas árvores pode ser computacionalmente intensa, exigindo mais recursos em comparação com modelos individuais. "Em conjuntos de dados muito grandes, a demanda de memória e tempo de processamento pode ser significativa" (Hastie; Robert; Jerome, 2009).

ii) **Menor Interpretabilidade**: Embora as árvores individuais sejam interpretáveis, o modelo global de *Random Forest* pode ser desafiador de interpretar como um todo, especialmente em casos de alto número de variáveis e árvores (Breiman, 2001).

O modelo *Random Forest* é amplamente adotado em áreas como medicina, finanças e agricultura de precisão. Segundo Breiman (2001), "sua robustez e precisão o tornam ideal para problemas que demandam previsões consistentes e confiáveis em cenários complexos".

2.7.4 Comparação entre os modelos

A comparação entre modelos convencionais e métodos de Aprendizado de Máquina é particularmente relevante para análises que envolvem complexidade e não linearidade, como é o caso desse estudo. Em contextos de previsão e classificação, essa comparação destaca as diferenças entre abordagens interpretativas e flexíveis. Modelos convencionais, incluindo técnicas estatísticas clássicas como a Regressão não linear, oferecem uma estrutura interpretativa mais rígida e permitem que se compreendam diretamente as relações entre as variáveis, ainda que assumam uma forma específica de não linearidade. Hastie, Robert e Jerome (2009) apontam que os modelos estatísticos clássicos oferecem uma estrutura rigorosa para a análise de dados, permitindo interpretações diretas dos coeficientes associados às variáveis.complexos, o que é útil para manter uma compreensão detalhada dos parâmetros, mesmo em modelos mais complexos.

Entretanto, esses modelos apresentam limitações, pois tendem a ser menos flexíveis e menos eficazes ao lidar com padrões não lineares complexos. Como destacado por James et al. (2013), "os modelos lineares convencionais podem não capturar adequadamente as complexidades presentes em muitos conjuntos de dados do mundo real".

Por outro lado, o Aprendizado de Máquina abrange uma ampla gama de algoritmos que podem aprender a partir de dados de maneira autônoma, permitindo uma modelagem mais robusta de relações não lineares. Murphy (2012) observa que "o Aprendizado de Máquina se destaca em tarefas que exigem a identificação de padrões complexos em grandes volumes de dados". Essa capacidade de adaptação e precisão torna os métodos de aprendizado de máquina especialmente valiosos em áreas como reconhecimento de imagem e análise preditiva.

Contudo, o uso de algoritmos de Aprendizado de Máquina também apresenta desafios, como a complexidade e a menor interpretabilidade. Breiman (2001) aponta que "modelos complexos, como redes neurais e ensembles, podem ser considerados caixas-pretas, dificultando a compreensão das decisões tomadas pelo modelo". Além disso, esses métodos geralmente requerem grandes quantidades de dados para treinamento eficaz, o que pode não estar disponível em todas as situações.

Em suma, a escolha entre modelos convencionais e métodos de Aprendizado de Máquina deve ser orientada pelo contexto do problema, pela natureza dos dados disponíveis e pelos objetivos da análise. Uma abordagem híbrida que combine as forças de ambos os tipos de modelos pode muitas vezes oferecer uma solução mais robusta e eficaz.

3 MATERIAL E MÉTODOS

Para a realização deste estudo, foi utilizado um banco de dados proveniente do curso "Modelos de Regressão Não Linear", conduzido por Zeviani, Ribeiro Júnior e Bonat (2013) em Curitiba-PR, Brasil, que utilizou o modelo de *Gompetz*, *Logístico* e o *Weibull* para realização da análise. Este conjunto de dados inclui informações sobre o crescimento dos frutos da goiabeira da variedade "Pedro Sato", com variáveis como peso, volume e comprimento longitudinal, que são essenciais para a análise de regressão.

As análises foram centradas na aplicação de modelos de Regressão não linear, que são particularmente adequados para descrever relações complexas entre variáveis biológicas. Os modelos clássicos foram aplicados, sendo que o modelo Logístico, de *Von Bertalanffy* e de *Richards* permitem representar o crescimento do fruto em função do comprimento longitudinal.

Tabela 3.1 – Modelos Clássicos de Regressão não-linear.

Modelos	Função
Logístico	$f(x) = \frac{\alpha}{1 + e^{(\beta - \gamma x)}}$
Von Bertalanffy	$f(x) = \alpha[1 - e^{-\gamma^{(x-\delta)}}]$
Richards	$f(x) = \frac{\alpha}{1 + e^{(\beta - \gamma x)^{1/\delta}}}$
Gompertz	$f(x) = \alpha e^{-e^{(\beta - \gamma x)}}$
Weibull	$f(x) = \alpha - \beta e^{-\gamma x^{\delta}}$
Morgan-Mercer-Flodin (MMF)	$f(x) = \frac{\beta\gamma + \alpha x^{\delta}}{\gamma + x^{\delta}}$

Fonte:(Seber; Lee, 2003)

Além disso, os modelos de Aprendizado de Máquina *Random Forest*, *Árvores de Decisão* e Máquina de Vetores de Suporte (SVM) foram implementado por ser especialmente eficaz em lidar com dados de alta dimensionalidade e pode capturar interações complexas entre as variáveis independentes e a variável de resposta.

Para avaliar o desempenho dos modelos, foram utilizadas métricas como a Raíz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2). O RMSE fornece uma medida da precisão das previsões ao indicar a média dos erros quadráticos, enquanto o R^2 quantifica a proporção da variância da variável dependente que é explicada pelo modelo. Essas métricas são essenciais para comparar a eficácia dos modelos ajustados e determinar qual abordagem oferece melhores previsões em relação aos dados observados.

4 RESULTADOS E DISCUSSÕES

A análise descritiva apresentadas, a seguir, fornece um panorama geral dos dados coletados. Esses valores incluem medidas de posição, que facilitam a compreensão da tendência central dos dados. Esses resumos estatísticos são fundamentais para orientar as próximas etapas da análise. A Tabela 4.1 apresenta a análise descritiva de cada variável.

Tabela 4.1 – Análise descritiva de características da Goiaba.

Estatísticas	**Comprimento Longitudinal**	**Peso**	**Volume**
Mínimo	29,15	6,27	6,0
1º Quartil	42,41	19,66	20,0
Mediana	67,33	89,56	95,0
Média	62,04	102,09	111,3
3º Quartil	81,31	178,21	200,0
Máximo	101,07	304,11	310,0

Fonte: Elaborado pelo autor, (2024).

Conforme observado na Tabela 4.1, o comprimento longitudinal apresenta um valor mínimo de 29,15 mm e máximo de 101,0 mm, com uma média de 62,04 mm. Essa distribuição sugere uma moderada variação nos comprimentos registrados, com a maioria dos valores concentrados entre 42,41 mm e 81,31 mm, correspondendo ao intervalo entre o primeiro e o terceiro quartil.

Para o peso, os valores variam amplamente, de um mínimo de 6,27g até um máximo de 304,11 g, com uma média de 102,09 g. A mediana de 89.56 g, significativamente abaixo da média, sugere a presença de valores elevados que podem estar influenciando a média, indicando uma distribuição possivelmente assimétrica.

O volume também mostra uma ampla amplitude, com valores variando entre 6,0 mm^3 e 310,0 mm^3, e uma média de 111,3 mm^3. Assim como o peso, a mediana de 95,0 mm^3, menor que a média, sugere que alguns valores altos estão presentes, o que contribui para aumentar a média geral.

Essas descrições ajudam a identificar tendências e possíveis assimetrias nos dados, permitindo ajustes e interpretações mais precisas nas análises futuras.

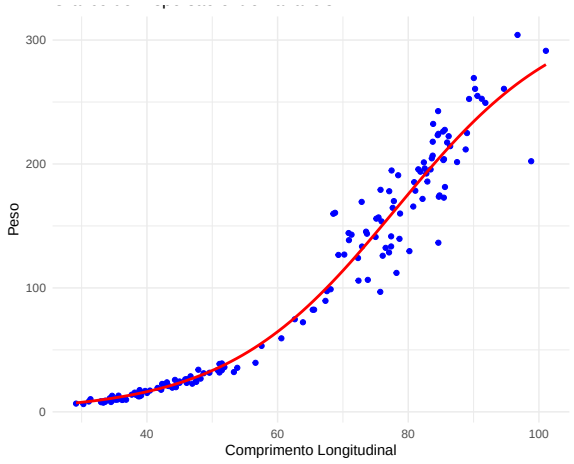
4.1 Modelo não Linear

Para os modelos não lineares, foram utilizados três modelos principais: o modelo Logístico, o modelo de *Von Bertalanffy* e o modelo de *Richards*. Cada um deles aplicado com o objetivo de capturar diferentes aspectos da relação entre o comprimento longitudinal, peso e volume, buscando uma representação fiel das dinâmicas observadas nos dados.

Além da aplicação individual dos modelos Logístico, de *Von Bertalanffy* e de *Richards*, realizou-se uma comparação entre eles para identificar o ajuste mais adequado aos dados observados. Essa comparação permite avaliar qual modelo captura melhor as nuances do crescimento da goiaba representado nas variáveis long, peso e volume, considerando características como o padrão de crescimento e o comportamento assimtótico.

Nas Figuras 4.1 e 4.2, será possivel observar a dispersão dos dados em termos de peso e volume de acordo com a longitude, respectivamente. Ambos os gráficos fornecem uma visão detalhada do comportamento dessas variáveis em função do comprimento, revelando uma relação de crescimento que pode ser modelada adequadamente por meio de regressões não-lineares.

Figura 4.1 – Gráfico de dispersão para o Peso e o Comprimento Longitudinal do fruto da goiabeira.

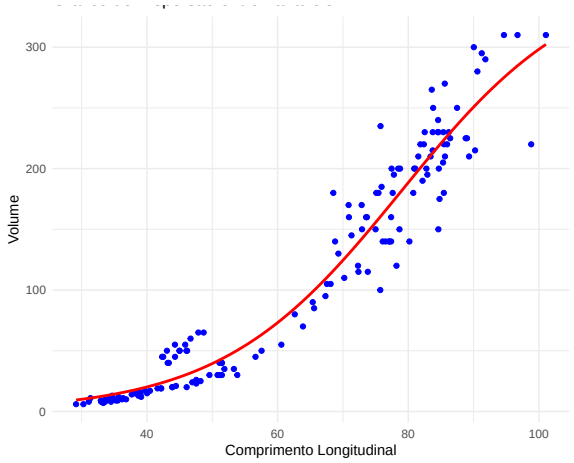


Fonte: Elaborado pelo autor, (2024).

No gráfico da Figura 4.1, observa-se um padrão de dispersão entre o comprimento longitudinal do fruto e o peso correspondente. À medida que o comprimento do fruto aumenta, há uma tendência clara de aumento do peso. Esse comportamento indica uma relação positiva entre as duas variáveis: frutos mais longos tendem a ser mais pesados.

A dispersão dos dados respresenta bem a tendência de crescimento do peso em relação ao comprimento longitudinal, mostrando uma trajetória de crescimento que se estabiliza à medida que o comprimento do fruto atinge valores maiores. Esse comportamento pode ser interpretado como um sinal de saturação, em que o peso do fruto atinge um limite à medida que o comprimento continua a aumentar, o que é característico de processos de crescimento biológico em plantas e frutos.

Figura 4.2 – Gráfico de dispersão para o Volume e o Comprimento Longitudinal do fruto da goiabeira.



Fonte: Elaborado pelo autor, (2024).

O gráfico da Figura 4.2, por sua vez, apresenta a relação entre o comprimento longitudinal e o volume do fruto da goiabeira. Novamente, é possível observar uma relação positiva entre essas variáveis, em que o volume do fruto aumenta proporcionalmente ao comprimento. Esse aumento no volume com o crescimento em comprimento indica que frutos mais longos também têm maior volume, seguindo uma tendência semelhante à observada entre o peso e o comprimento.

Assim como no gráfico em 4.1 a curva modela bem os dados, e também sugere uma saturação, no qual o volume tende a estabilizar à medida que o comprimento do fruto aumenta, o que também é consistente com os padrões biológicos de crescimento.

A partir da análise detalhada desses gráficos, fica evidente que há uma relação positiva e não-linear entre o comprimento longitudinal do fruto da goiabeira e o peso e o volume. Em ambos os casos, a curva capturando a tendência geral de crescimento e as características de saturação. Esse comportamento sugere que, à medida que o fruto atinge um certo comprimento, tanto o peso quanto o volume começam a estabilizar, o que é comum em modelos biológicos de crescimento, em que fatores limitantes (como capacidade de suporte dos tecidos do fruto) podem levar a uma estabilização.

A escolha de modelos não-lineares, como o modelo Logístico, de *Von Bertalanffy* e de *Richards*, se justifica por representar bem a complexidade do fenômeno observado e pela adequação aos dados experimentais apresentados. Dessa forma, esses modelos oferecem uma ferramenta eficaz para prever o comportamento do peso e do volume com base no comprimento longitudinal dos frutos, contribuindo para uma compreensão mais aprofundada do crescimento da goiabeira e otimizando a análise do potencial produtivo da planta. Na Tabela 4.2 são apresentados os valores calculados para as estimativas dos

parâmetros dos modelos Logístico, *Von Bertalanffy* e *Richards*, com seus respectivos erros padrão, valores t e intervalos de confiança de 97,5%.

Tabela 4.2 – Valores de estimativa dos parâmetros; erro padrão de estimativa, valor-p e intervalo de confiança para os modelos.

Modelo Não Linear	Parâmetro	Estimativa	Erro Padrão	Pr(>| t |)	Intervalo de Confiança
Logístico (Peso)	α	169,9413	6,0122	<0,01 ***	[158,16; 181,73]
	β	0,1401	0,0267	<0,01 ***	[0,0877; 0,1925]
	γ	50,0000	1,5714	<0,01 ***	[46,92; 53,08]
Logístico (Volume)	α	184,5215	7,0653	<0,01 ***	[170,67; 198,37]
	β	0,1272	0,0241	<0,01 ***	[0,0799; 0,1745]
	γ	50,0000	1,7190	<0,01 ***	[46,63; 53,37]
Von Bertalanffy (Peso)	α	200,0000	9,4518	<0,01 ***	[181,47; 218,53]
	β	0,0643	0,0058	<0,01 ***	[0,0531; 0,0756]
	γ	35,6947	1,0229	<0,01 ***	[33,69; 37,70]
Von Bertalanffy (Volume)	α	300,0000	29,2298	<0,01 ***	[242,71; 357,29]
	β	0,0419	0,0061	<0,01 ***	[0,0299; 0,0540]
	γ	33,2020	2,3226	<0,01 ***	[28,65; 37,75]
Richards (Peso)	α	192,1793	16,3567	<0,01 ***	[160,12; 224,24]
	β	0,0735	0,0328	0,0259 **	[0,0095; 0,1375]
	γ	50,0000	4,5373	<0,01 ***	[41,11; 58,89]
Richards (Volume)	α	211,7684	19,5893	<0,01 ***	[173,37; 250,16]
	β	0,0701	0,0325	0,0325 **	[0,0064; 0,1339]
	γ	50,0000	4,8347	<0,01 ***	[40,52; 59,48]

Fonte: Elaborado pelo autor, (2024).

A Tabela 4.2 resume as estimativas dos parâmetros dos modelos Logístico, *Von Bertalanffy* e *Richards* aplicados às variáveis de peso e volume, respectivamente. Cada modelo foi ajustado de forma a capturar o comportamento de crescimento das variáveis analisadas, e os parâmetros de cada um refletem as especificidades de suas respectivas estruturas matemáticas.

Para o modelo Logístico, os valores das estimativas sugerem um crescimento sigmoidal para as variáveis de peso e volume, com o parâmetro α representando o valor assintótico máximo que a variável pode alcançar. Os baixos valores de erro padrão e os valores p significativos (<0,01) indicam que os parâmetros são estatisticamente relevantes, refletindo uma curva de ajuste robusta, característica que, segundo Zeviani, Ribeiro Júnior e Bonat (2013), torna este modelo adequado para descrever processos de crescimento em organismos vivos.

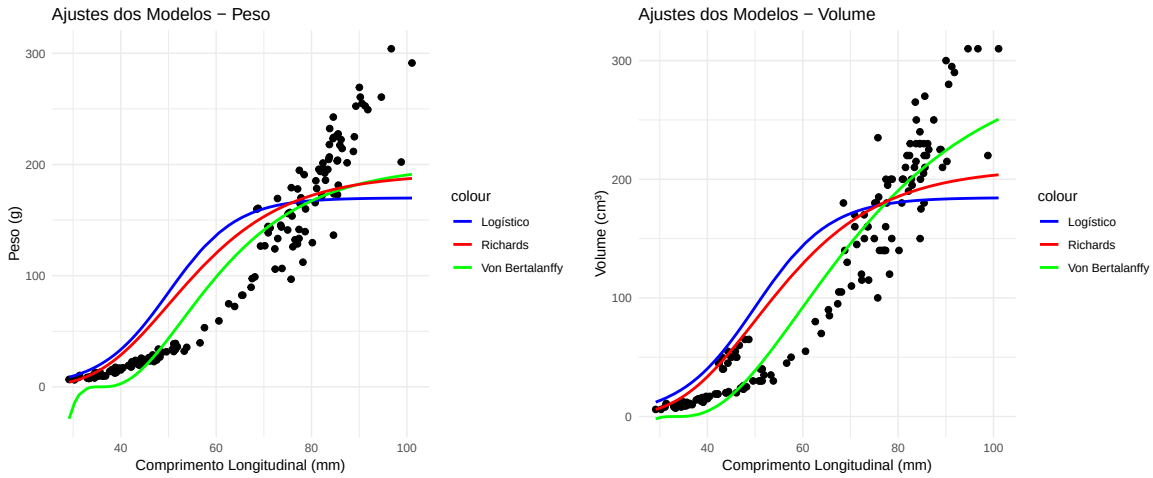
No caso do modelo *Von Bertalanffy*, as estimativas mostram valores de α superiores ao modelo Logístico, particularmente para a variável volume, sugerindo um comportamento

de crescimento assintótico mais elevado. A significância estatística dos parâmetros, aliada aos baixos erros padrão, sugere que esse modelo pode capturar bem a dinâmica de crescimento gradual e assintótica observada nos dados, especialmente para processos de crescimento fisiológico (Zeviani; Ribeiro Júnior; Bonat, 2013).

Para o modelo *Richards* os parâmetros α, β e γ foram estimados para descrever o padrão de crescimento das variáveis, oferecendo maior flexibilidade no ajuste da curva. Os valores de intervalo de confiança para os parâmetros α e γ são amplos, o que é desfavorável, pois indica que o modelo está acomodando uma variação excessiva nos dados observados, sugerindo menor precisão nas previsões.

Em resumo, as estimativas fornecidas pelos três modelos indicam que tanto o modelo Logístico quanto o de *Von Bertalanffy* são adequados para descrever o crescimento das variáveis de interesse, com o modelo *Richards* adicionando flexibilidade extra, embora com maior incerteza em alguns parâmetros. Esses resultados sugerem que cada modelo oferece informações distintas sobre a dinâmica de crescimento dos dados analisados. A seguir, serão apresentados nas Figuras 4.3, os gráficos dos ajustes dos três parâmetros, verificando a relação entre as variáveis peso e volume com o comprimento longitudinal da goiaba.

Figura 4.3 – Gráfico de ajuste para os modelos Logístico, *Von Bertalanffly* e *Richards* para a variável Peso e Volume.



Fonte: Elaborado pelo autor, (2024).

A visualização gráfica permite observar que o modelo de *Von Bertalanffly* tem um ajuste ligeiramente melhor aos dados em relação aos demais modelos. Para confirmar esse melhor ajuste aos dados, foram utilizados os critérios apresentados na Tabela 4.3.

Tabela 4.3 – Critérios de Informação de Akaike (AIC), Bayesiano de Schwarz (BIC) e a Raíz do Erro Quadrático Médio (RMSE) para os modelos ajustados.

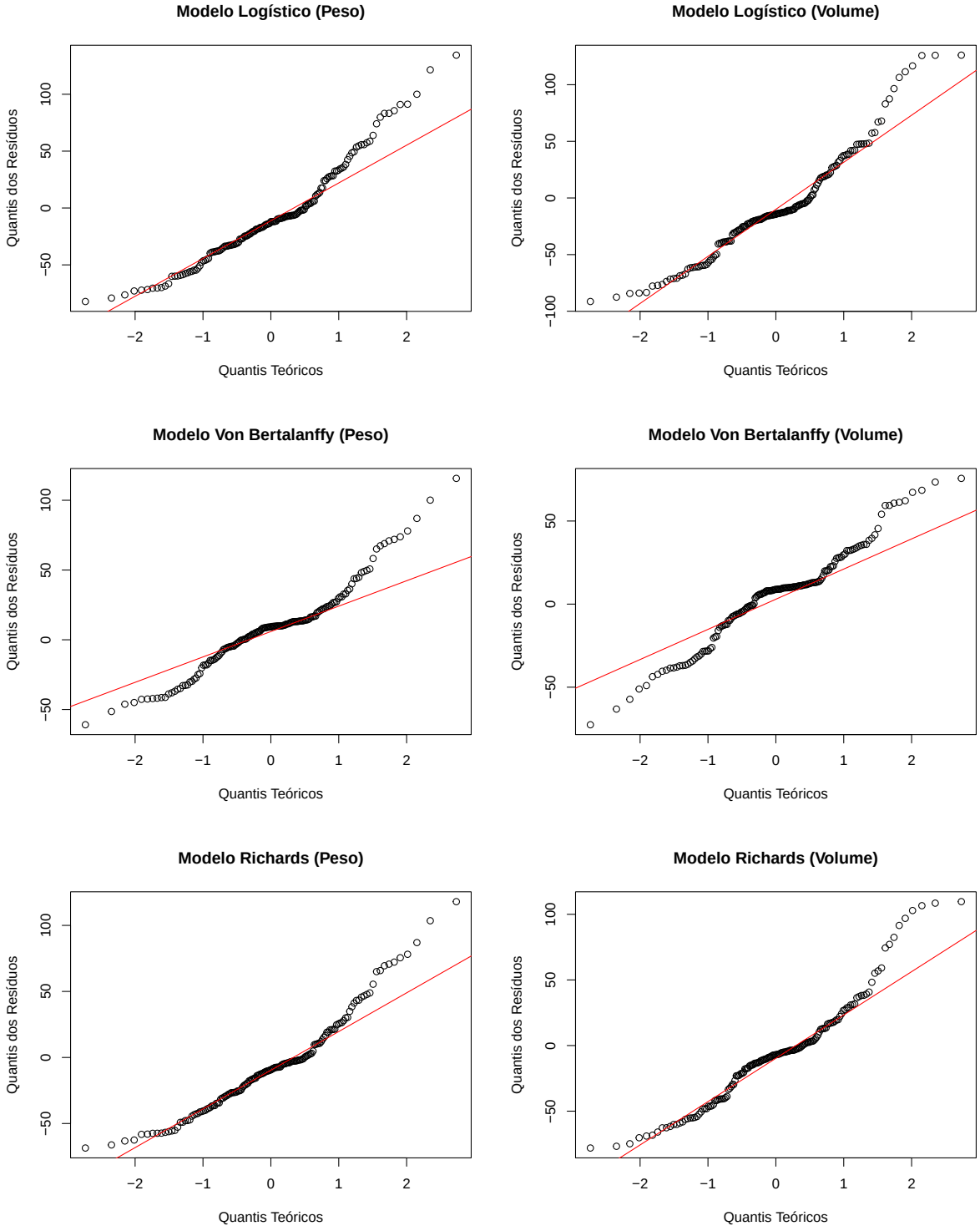
Modelo Não Linear	AIC	BIC	RMSE
Logístico (Peso)	1653,1780	1665,4540	42,7162
Logístico (Volume)	1678,1130	1690,3880	46,2004
Von Bertalanffy (Peso)	1546,5820	1558,8580	30,5502
Von Bertalanffy (Volume)	1517,3620	1529,6380	27,8681
Richards (Peso)	1599,0090	1614,3530	35,7999
Richards (Volume)	1627,2910	1642,6350	39,1297

Fonte: Elaborado pelo autor, (2024)

Com a apresentação dos critérios AIC, BIC e RMSE para os modelos não lineares utilizados na análise do crescimento do fruto da goiabeira "Pedro Sato", pode-se observar resultados significativos que refletem a performance de cada modelo em relação às variáveis peso e volume. É possível observar que o modelo *Von Bertalanffly* se destacou significativamente comparado aos demais modelos. O modelo *Richards* apresentou resultados intermediários. E o modelo Logístico apresentou resultados que deixaram a desejar, ficando como o pior modelo entre os três. O modelo *Von Bertalanffly* se destacou como o mais eficaz, tanto em termos de critérios de informação (AIC e BIC) quanto de precisão (RMSE) (Tabela 4.3).

Nas Figuras 4.4 são apresentados os gráficos de normalidade dos resíduos dos modelo. O gráfico de normalidade dos resíduos para a variável peso mostra um alinhamento razoável dos pontos com a linha de referência na região central, sugerindo uma aproximação à normalidade para grande parte dos resíduos. Contudo, há um desvio nas caudas, com pontos que se afastam da linha reta. Esse comportamento indica que, embora o modelo Logístico capture o padrão geral do crescimento em peso do fruto da goiabeira "Pedro Sato", existem valores extremos que podem refletir variações naturais no crescimento dos frutos ou limitações do modelo em ajustar completamente esses dados. Já para a variável Volume, o gráfico QQ exibe um padrão semelhante. A maior parte dos pontos segue a linha de referência na região central, mas desvios nas extremidades sugerem que o modelo Logístico tem dificuldades em capturar as variações mais extremas do volume dos frutos. Esses desvios podem indicar uma variação no desenvolvimento volumétrico das goiabas "Pedro Sato" que não é totalmente explicada pelo modelo Logístico.

Figura 4.4 – Gráfico de Normalidade dos Resíduos dos Modelos Não Líneares para o Peso e o Volume.



Fonte: Elaborado pelo autor, (2024).

No gráfico de normalidade dos resíduos para Peso, os pontos seguem a linha de referência na maior parte do gráfico, indicando incialmente que a normalidade dos resíduos é razoavelmente atendida. Contudo, pequenos desvios nas caudas podem sugerir que o

modelo *Von Bertalanffy* ajusta bem o crescimento em peso da goiaba "Pedro Sato", mas ainda apresenta alguma limitação em relação aos valores extremos, o que pode ser associado a fatores ambientais ou genéticos que afetam o desenvolvimento do fruto. Para Volume, observa-se um comportamento semelhante ao Peso, com os resíduos alinhados na maior parte do gráfico, mas com leve afastamento nas extremidades. Esses desvios podem indicar variações no crescimento volumétrico que o modelo *Von Bertalanffy* não capta completamente, possivelmente relacionadas a aspectos de irrigação, fertilização ou condições climáticas que impactam o volume dos frutos.

Para o peso, os pontos no gráfico de normalidade dos resíduos estão bem alinhados com a linha de referência, indicando uma aproximação à normalidade superior em relação aos modelos anteriores. Isso sugere que o modelo *Richards* oferece um ajuste mais fiel ao crescimento em peso do fruto "Pedro Sato", capturando melhor as variações nos dados. Para o volume, o modelo *Richards* também mostra um alinhamento consistente dos resíduos com a linha de referência. Pequenos desvios nas extremidades indicam que, embora haja algumas variações não capturadas, o modelo *Richards* é adequado para descrever o crescimento em volume do fruto "Pedro Sato".

A análise dos gráficos de normalidade dos resíduos para os resíduos dos modelos Logístico, *Von Bertalanffly* e *Richards* indica que o modelo *Richards* é o mais apropriado para descrever o crescimento do fruto da goiabeira "Pedro Sato", tanto para o peso quanto para o volume, uma vez que apresenta resíduos mais próximos de uma distribuição Normal. Esse resultado sugere que o modelo *Richards* é particularmente útil para entender o desenvolvimento desses frutos, fornecendo uma ferramenta útil para otimizar práticas agrícolas e melhorar a produtividade da variedade "Pedro Sato".

Tabela 4.4 – Teste de normalidade de Shapiro-Wilk.

Modelo	**Variável**	**Estatística W**	**Valor-*p***	
Logístico	Peso	0,9495	<0,01	***
Logístico	Volume	0,9462	<0,01	***
Von Bertalanffy	Peso	0,9534	<0,01	***
Von Bertalanffy	Volume	0,9699	<0,01	***
Richards	Peso	0,9498	<0,01	***
Richards	Volume	0,9479	<0,01	***

Fonte: Elaborado pelo autor, (2024).

Com base nos resultados dos testes de normalidade para os resíduos, conclui-se que todos os modelos (Logístico, *Von Bertalanffy* e *Richards*) apresentam resíduos que não seguem uma distribuição normal para ambas as variáveis Peso e Volume. Esse comportamento não-normal dos resíduos pode indicar que os modelos ajustados não capturam

perfeitamente a variabilidade dos dados, o que pode ter implicações na precisão das previsões e na inferência estatística.

4.2 Aplicação do *Machine Learning*

Para compreender a eficiência dos diferentes modelos de Aprendizado de Máquina aplicados na estimativa de características do fruto da goiabeira "Pedro Sato", foram realizados testes com três abordagens distintas: Árvore de Decisão, *Random Forest* e Máquina de Vetores de Suporte (SVM). São apresentados, na Tabela 4.5 os resultados obtidos para o RMSE para o peso e o volume dos frutos.

Tabela 4.5 – Raiz do Erro Quadrático Médio (RMSE) dos Modelos de Aprendizado de Máquina para o Peso e o Volume

	Peso	**Volume**
Modelo	**RMSE**	**RMSE**
Árvore de Decisão	14,1929	23,6232
Random Forest	7,6799	14,5256
SVM	13,8048	15,7152

Fonte: Elaborado pelo autor, (2024)

Analisando os dados da Tabela 4.5 é possível observar que o modelo *Random Forest* apresentou o menor RMSE tanto para a predição do peso (7.6799) quanto para a predição do volume (14.5256), indicando uma maior precisão em comparação aos outros modelos testados. O modelo da Máquina de Vetores de Suporte (SVM) obteve um desempenho intermediário em ambas as predições, enquanto a Árvore de Decisão demonstrou o maior erro médio quadrático, o que sugere menor acurácia em estimar as características dos frutos. Esses resultados reforçam que, para o contexto específico da análise de frutos da goiabeira "Pedro Sato", a *Random Forest* se destaca como a abordagem mais eficiente, possivelmente devido à sua capacidade de capturar interações complexas entre variáveis e reduzir o viés e a variância do modelo.

4.3 Comparação dos Modelos

Para avaliar o desempenho dos modelos de Aprendizado de Máquina aplicados à predição das características dos frutos da goiabeira "Pedro Sato", foram calculados os valores de RMSE para os modelos *Random Forest* e Von Bertalanffy, tanto para o peso quanto para o volume dos frutos. A seguir, são apresentados os resultados de RMSE para cada modelo.

Os resultados da Tabela 4.6 indicam que, para a predição do peso dos frutos, o modelo *Random Forest* apresentou um RMSE de 7.6799, superando o modelo *Von Bertalanffy*, que obteve um RMSE de 30.55016. Da mesma forma, para a predição do volume, o modelo

Random Forest também mostrou um desempenho superior, com um RMSE de 14.5256 comparado a 27.86812 do modelo *Von Bertalanffy*. Esses resultados sugerem que, em ambos os casos, o modelo *Random Forest* é mais preciso na estimativa das características dos frutos da goiabeira "Pedro Sato", destacando-se como a melhor opção entre os modelos avaliados.

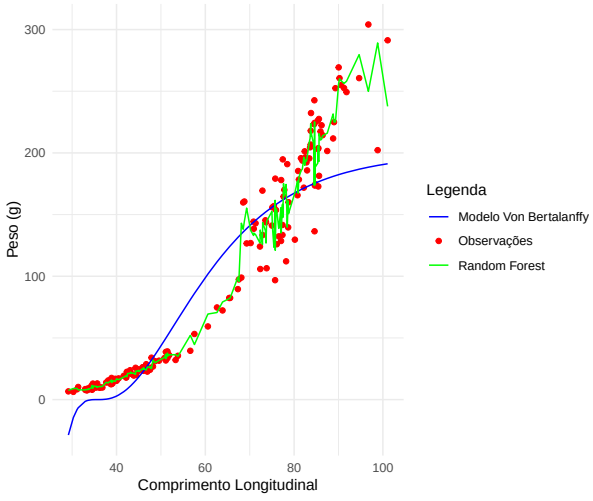
Tabela 4.6 – Raiz do Erro Quadrático Médio (RMSE) para comparação do Modelo de *Von Bertalanffy* e *Random Forest* para o Peso e o Volume

	Peso	**Volume**
Modelo	**RMSE**	**RMSE**
Random Forest	7,6799	14,5256
Von Bertalanffy	30,5502	27,8681

Fonte: Elaborado pelo autor, (2024)

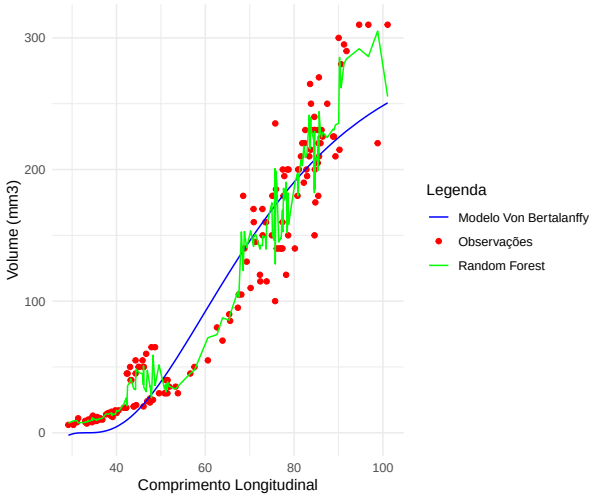
Para uma melhor compreensão do ajuste dos modelos de *Von Bertalanffy* e *Random Forest* sobre as variáveis Peso e Volume em função do comprimento Longitudinal dos frutos, foi realizada uma análise comparativa das curvas geradas por cada modelo em relação aos dados observados. A escolha do modelo de *Von Bertalanffy* deve-se à sua característica de crescimento não linear, adequada para representar o crescimento de organismos, enquanto o modelo *Random Forest*, sendo um algoritmo de Aprendizado de Máquina, pode capturar relações complexas e não lineares sem pressupor uma forma funcional específica.

Figura 4.5 – Comparação das curvas dos modelos de *Von Bertalanffy* e *Random Forest* para o Peso



Fonte: Elaborado pelo autor, (2024)

Figura 4.6 – Comparação das curvas dos modelos de *Von Bertalanffy* e *Random Forest* para o Volume



Fonte: Elaborado pelo autor, (2024)

Observa-se nas Figuras 4.5 e 4.6 que o modelo de *Von Bertalanffy* apresenta um ajuste suave e contínuo, que capta padrão de crescimento dos frutos, especialmente em intervalos intermediários de comprimento longitudinal. No entanto, em certos pontos, principalmente em comprimentos maiores, o modelo *Random Forest* parece acompanhar melhor as flutuações presentes nos dados observados, capturando variações que o modelo de *Von Bertalanffy* não representa tão bem.

Essa comparação evidencia que, enquanto o modelo de *Von Bertalanffy* é capaz de descrever o padrão geral de crescimento do fruto sem captar de forma eficaz, já o modelo *Random Forest* demonstra maior flexibilidade ao capturar variações locais nas goiabas, captando melhor o padrão de crescimento. Esse comportamento sugere que a escolha do modelo ideal é o *Random Forest*, pois ele é preciso em pontos específicos do comprimento longitudinal.

Os resultados obtidos neste trabalho mostram semelhanças e diferenças em relação aos estudos de Conceição (2022) e Carvalho Júnior et al. (2016). Em ambos os casos, foram analisados modelos estatísticos e técnicas de Aprendizado de Máquina para avaliar fenômenos complexos. No estudo de Carvalho Júnior et al. (2016), foram comparados modelos de Regressão linear múltipla (RLM) e *Random Forest* (RF), destacando a importância de variáveis como carbono orgânico e frações granulométricas para a estimativa da densidade do solo. Já em Conceição (2022), o foco foi na previsão da inflação utilizando *Random Forest* e métodos tradicionais, como ARIMA, observando-se que o *Random Forest* apresentou desempenho superior devido à sua capacidade de lidar com não linearidades e selecionar variáveis relevantes.

No presente trabalho, a comparação entre modelos não lineares (Logístico, *Von Ber-*

talanffy e *Richards*) e algoritmos de Aprendizado de Máquina também evidencia a importância de diferentes abordagens modelísticas para melhorar a precisão preditiva. Semelhante aos resultados de Conceição (2022), o *Random Forest* destacou-se pela capacidade de ajustar melhor os dados, conforme métricas de validação como RMSE. Por outro lado, ao contrário do foco dos artigos citados na separação detalhada entre R^2 de ajuste e validação, este estudo priorizou a análise do ajuste global.

Adicionalmente, os resultados mostram que o presente estudo compartilha semelhanças metodológicas significativas com os trabalhos citados, sobretudo no uso de métricas robustas para avaliação de modelos e na seleção criteriosa de variáveis preditoras. Esse alinhamento metodológico reforça a importância de combinar métodos tradicionais e modernos para capturar relações complexas entre as variáveis. Conclui-se, como apontado por Conceição (2022) e Carvalho Júnior et al. (2016), que essa abordagem integrada oferece uma solução robusta para modelagem preditiva, seja na previsão macroeconômica, no estudo de propriedades físicas do solo ou em aplicações biológicas, como neste trabalho, contribuindo para a redução de incertezas e a melhora na precisão das estimativas.

5 CONCLUSÃO

Este trabalho teve como objetivo aplicar a Regressão não linear multiresposta e comparar modelos estatísticos e de Aprendizado de Máquina para descrever o crescimento dos frutos da goiabeira "Pedro Sato". Foram utilizados modelos clássicos de Regressão não linear, como os modelos Logístico, *Von Bertalanffy* e *Richards*, além de algoritmos de Aprendizado de Máquina, incluindo *Random Forest*, Árvore de Decisão e Máquina de Vetores de Suporte (SVM). A avaliação dos modelos foi feita com base no AIC, BIC e RMSE, permitindo identificar os modelos que melhor se ajustaram aos dados e forneceram as previsões mais precisas.

Embora o modelo de *Von Bertalanffy* tenha sido adequado para capturar o padrão geral de crescimento, o *Random Forest* superou os modelos não lineares, apresentando um RMSE significativamente menor. Sua flexibilidade em capturar interações complexas e padrões não lineares, sem a necessidade de uma função paramétrica, possibilitou uma descrição precisa das variações locais nos dados, principalmente em pontos específicos de crescimento.

Os resultados indicam que o *Random Forest* é o modelo mais adequado para descrever o crescimento dos frutos, permitindo previsões precisas e ajustes específicos no manejo agrícola. Esse modelo pode ser útil para a definição do ponto ideal de colheita, otimização de recursos e redução de perdas, maximizando a eficiência e a lucratividade. Sua aplicação pode melhorar o planejamento das etapas de colheita e comercialização, além de contribuir para a qualidade final da produção.

Apesar dos resultados positivos, algumas limitações foram identificadas, como a natureza destrutiva das amostras, que restringiu a análise longitudinal. Recomenda-se o uso de tecnologias não destrutivas, como o sensoriamento remoto, e a inclusão de variáveis ambientais, como temperatura e umidade, para uma análise mais abrangente e precisa do crescimento dos frutos. Além disso, a integração com redes neurais pode melhorar ainda mais a acurácia das previsões em dados complexos e de grande volume.

Em resumo, o trabalho demonstra a aplicabilidade e eficiência dos modelos de Regressão não linear multiresposta e Aprendizado de Máquina no estudo do crescimento dos frutos. A escolha do modelo ideal, *Random Forest*, mostrou-se eficaz para capturar detalhes com precisão. Este estudo reforça a importância de métodos avançados de modelagem e abre caminhos para futuras pesquisas, contribuindo para a agricultura e o avanço da ciência na análise do crescimento vegetal.

REFERÊNCIAS

AITKEN, A. C. On Least Squares and Linear Combinations of Observations. **Proceedings of the Royal Society of Edinburgh**, v. 56, p. 56–75, 1936. DOI: ⟨10.1017/S0370164600011788⟩.

AKAIKE, H. A New Look at the Statistical Model Identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974. ISSN 0018-9286. DOI: ⟨10.1109/TAC.1974.1100705⟩.

ALPAYDIN, E. **Introduction to Machine Learning**. [S.l.]: MIT Press, 2020. ISBN 978-0262039406.

BATES, D. M.; WATTS, D. G. **Nonlinear Regression Analysis and Its Applications**. [S.l.]: John Wiley & Sons, 1988. P. 1–365. ISBN 978-0471816434.

BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. DOI: ⟨10.1023/A:1010933404324⟩.

BREIMAN, Leo; FRIEDMAN, Jerome; OLSHEN, Richard; STONE, Charles. **Classification and Regression Trees**. Monterey, CA: Wadsworth e Brooks/Cole Advanced Books Software, 1984.

BREIMAN, Leo; FRIEDMAN, Jerome; OLSHEN, Richard A.; STONE, Charles J. **Classification and Regression Trees**. [S.l.]: Wadsworth Brooks/Cole, 1986.

BROYDEN, C. G.; FLETCHER, R.; GOLDFARB, D.; SHANNO, D. **Numerical Methods for Unconstrained Optimization**. New York: Wiley-Interscience, 1970.

BURDEN, Richard L.; FAIRES, J. Douglas. **Numerical Analysis**. 10th. Boston: Cengage Learning, 2016.

BURNHAM, Kenneth P.; R., Anderson David. **Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach**. [S.l.]: Springer, 2002.

CARVALHO JÚNIOR, Waldir de; FILHO, Braz Calderano; SILVA CHAGAS, César da; BHERING, Silvio Barge; PEREIRA, Nilson Rendeiro;
PINHEIRO, Helena Saraiva Koenow. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. **Pesquisa Agropecuária Brasileira**, v. 51, n. 9, p. 1428–1437, 2016. DOI: ⟨10.1590/S0100-204X2016000900041⟩. Disponível em: ¡⟨http://dx.doi.org/10.1590/S0100-204X2016000900041⟩¿.

CASELLA, G.; BERGER, R. L. **Statistical Inference**. 2nd. [S.l.]: Duxbury, 2002. P. 1–660. ISBN 978-0534243128.

CONCEIÇÃO, Lia Souto Manhães da. **Previsão da inflação utilizando métodos tradicionais e aprendizado de máquina**. 2022. Dissertação de Mestrado – Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, Brasil.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. DOI: ⟨10.1007/BF00994018⟩.

DAVIDON, W. C. Variable Metric Method for Minimization. **SIAM Journal on Optimization**, v. 3, p. 1–10, 1959.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. 3rd. [S.l.]: Wiley-Interscience, 1998. P. 1–736. ISBN 978-0471170826.

FISHER, R. A. **Statistical Methods for Research Workers**. [S.l.]: Oliver e Boyd, 1925. P. 1–356.

_____. The Goodness of Fit. **Annals of Eugenics**, v. 1, p. 191–211, 1922. DOI: ⟨10.1111/j.1469-1809.1922.tb02148.x⟩.

FOGLIATTO, Flávio Sanson. Otimização de experimentos com variáveis de resposta descritas por perfis. **Pesquisa Operacional**, v. 28, n. 3, p. 679–688, 2008. DOI: ⟨10.1590/S0101-74382008000300010⟩.

GAUSS, C. F. **Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium**. [S.l.]: Perthes et Besser, 1809. P. 1–884. Classic work on celestial mechanics.

GOMPERTZ, Benjamin. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. **Philosophical Transactions of the Royal Society of London**, v. 115, p. 513–585, 1825. DOI: ⟨10.1098/rstl.1825.0026⟩.

GOODFELLOW, I. et al. Deep Learning. **MIT Press**, p. 1–775, 2016.

HANSEN, L. P. Large Sample Properties of Generalized Method of Moments Estimators. **Econometrica**, v. 50, n. 4, p. 1029–1054, 1982. ISSN 0012-9682. DOI: ⟨10.2307/1912775⟩.

HASTIE, Trevor; ROBERT, Tibshirani; JEROME, Friedman. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. New York: Springer, 2009. ISBN 978-0-387-84857-0.

HASTINGS, W. The Application of Regression Analysis in Data Analysis. **Journal of Statistical Methods**, v. 15, n. 2, p. 103–112, 2000.

HYNDMAN, Rob J.; B., Koehler Abel. Another look at measures of forecast accuracy. **International Journal of Forecasting**, Elsevier, v. 22, n. 4, p. 679–688, 2006.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An Introduction to Statistical Learning: with Applications in R**. New York: Springer, 2013.

LEGENDRE, A.-M. **Nouvelles Méthodes pour la Détermination des Orbites des Comètes**. [S.l.]: F. Didot, 1805. P. 1–280. Seminal work on least squares.

MORGAN, P. H.; MERCER, L. P.; FLODIN, N. W. General model for nutritional responses of higher organisms. **Proceedings of the National Academy of Sciences**, v. 72, n. 11, p. 4327–4331, 1975. DOI: ⟨10.1073/pnas.72.11.4327⟩.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [S.l.]: MIT Press, 2012. P. 1–1022. ISBN 978-0262018029.

NOCEDAL, J.; WRIGHT, S. J. **Numerical Optimization**. 2nd. [S.l.]: Springer, 2006. P. 1–664. ISBN 978-0387303031.

RATKOWSKY, D. A. **Handbook of Nonlinear Regression Models**. [S.l.]: Marcel Dekker, 1990. P. 1–241. ISBN 978-0824782710.

RICHARDS, F.J. A flexible growth function for empirical use. **Journal of Experimental Botany**, v. 10, n. 29, p. 290–301, 1959.

RUSSELL, Stuart; PETER, Norvig. **Artificial Intelligence: A Modern Approach**. 4th. London: Pearson, 2020.

SANTOS, F.J.; SILVA, A.B. Aplicação de modelos não lineares no crescimento de plantas frutíferas. **Revista Brasileira de Agricultura**, v. 88, n. 4, p. 234–245, 2013.

SCHÖLKOPF, Bernhard; SMOLA, Alexander J. **Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond**. [S.l.]: MIT Press, 2002.

SCHWARZ, Gideon E. Estimating the dimension of a model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

SEBER, G. A. F.; LEE, A. J. **Linear Regression Analysis**. 2nd. [S.l.]: Wiley, 2003. P. 1–582. ISBN 978-0471415408.

SEBER, G. A. F.; WILD, C. J. **Nonlinear Regression**. [S.l.]: Wiley, 1989. P. 1–768. ISBN 978-0471617604.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3-4, p. 591–611, 1965.

STIGLER, S. M. **The History of Statistics: The Measurement of Uncertainty before 1900**. [S.l.]: Harvard University Press, 1986. P. 1–432. ISBN 978-0674403413.

SUTTON, R. S.; G., Barto A. **Reinforcement Learning: An Introduction**. 2nd. [S.l.]: MIT Press, 2018. P. 1–525. ISBN 978-0262039246.

TEAM, R Core. **R: A Language and Environment for Statistical Computing**. [S.l.: s.n.], 2024. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: ¡⟨https://www.R-project.org/⟩¿.

VAPNIK, Vladimir. **Statistical Learning Theory**. New York: Wiley, 1998.

VON BERTALANFFY, Ludwig. Quantitative laws in metabolism and growth. **The Quarterly Review of Biology**, v. 32, p. 217–231, 1957.

WEIBULL, Waloddi. A Statistical Distribution Function of Wide Applicability. **Journal of Applied Mechanics**, v. 18, p. 293–297, 1951.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the Mean Absolute Error (MAE) over the Root Mean Squared Error (RMSE) in Assessing the Accuracy of Model-Simulated Daily Precipitation. **Climate Research**, v. 30, n. 3, p. 93–107, 2005.

ZEVIANI, W. M.; RIBEIRO JÚNIOR, P. J.; BONAT, W. H. **Curso: Modelos de Regressão Não Linear**. Curitiba, PR, Brazil: [s.n.], jul. 2013. Curso realizado de 22 a 26 de julho de 2013.