UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Marcos Vinício Cesario dos Santos

Análise dos Efeitos do ENOS sobre o Clima e a Produção Agrícola no Brasil: Um Estudo de Caso dos Estados de Goiás e Rio Grande do Sul

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Marcos Vinício Cesario dos Santos

Título do trabalho: Análise dos Efeitos do ENOS sobre o Clima e a Produção Agrícola no Brasil: Um Estudo de Caso dos Estados de Goiás e Rio Grande do Sul

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **David Henriques Da Matta**, **Professor do Magistério Superior**, em 02/12/2025, às 15:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Marcos Vinício Cesario Dos Santos**, **Discente**, em 02/12/2025, às 19:39, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Marcos Vinício Cesario dos Santos

Análise dos Efeitos do ENOS sobre o Clima e a Produção Agrícola no Brasil: Um Estudo de Caso dos Estados de Goiás e Rio Grande do Sul

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Goiás para aprovação no componente curricular TCC, como parte das exigências para a obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. David Henriques da Matta.

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Santos, Marcos Vinício Cesario dos
Análise dos Efeitos do ENOS sobre o Clima e a Produção Agrícola no Brasil [manuscrito] : Um Estudo de Caso dos Estados de Goiás e Rio Grande do Sul / Marcos Vinício Cesario dos Santos. - 2025.
47 f.: il.

Orientador: Prof. David Henriques da Matta.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Estatística, Goiânia, 2025.

Inclui siglas, abreviaturas, símbolos, gráfico, tabelas, lista de figuras, lista de tabelas.

1. El Niño–oscilação Sul (ENOS). 2. Análise de dados funcionais. 3. Ciclo de produção agrícola. 4. Random forest. 5. Valores SHAP. I. Matta, David Henriques da, orient. II. Título.

CDU 519.22

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos vinte e seis dias do mês de novembro do ano de 2025 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Análise dos Efeitos do ENOS sobre o Clima e a Produção Agrícola no Brasil: Um Estudo de Caso dos Estados de Goias e Rio Grande do Sul”, de autoria de Marcos Vinício Cesario dos Santos, do curso de Estatística, do Instituto de Matemática e Estatística da UFG. Os trabalhos foram instalados pelo Prof. Dr. David Henriques da Matta com a participação dos demais membros da Banca Examinadora: Luis Rodrigo Fernandes Baumann (IME/UFG), Mario Ernesto Piscoya Diaz (IME/UFG) e Ludmilla Ferreira Justino (Embrapa/GO). Após a apresentação, a banca examinadora realizou a arguição do estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 8,8 , tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Mario Ernesto Piscoya Diaz**, **Professor do Magistério Superior**, em 26/11/2025, às 18:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **David Henriques Da Matta**, **Professor do Magistério Superior**, em 26/11/2025, às 19:06, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Ludmilla Ferreira Justino**, **Usuário Externo**, em 27/11/2025, às 15:48, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Luis Rodrigo Fernandes Baumann**, **Professor do Magistério Superior**, em 01/12/2025, às 10:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5793891** e o código CRC **E99BF751**.

Referência: Processo nº 23070.059685/2025-39 SEI nº 5793891

Agradecimentos

Quero expressar minha profunda gratidão, primeiramente, aos meus pais, pelo apoio incondicional em todas as minhas decisões. Eles foram minha base e me mantiveram firme ao longo de todo este processo.

Agradeço também aos meus avós, que sempre me incentivaram a estudar. Em especial à minha avó Badia, que, mesmo com tantos netos, nunca se esquece de mim ao preparar aqueles deliciosos doces caseiros.

Aos demais familiares, meu sincero agradecimento pelo incentivo constante ao longo dos anos de graduação.

Ao meu orientador, sou grato pelas incontáveis contribuições que foram fundamentais para o meu crescimento pessoal e profissional.

Ao supervisor Alexandre, agradeço pelas inúmeras orientações e pela dedicação ao meu desenvolvimento, tanto pessoal quanto profissional.

E, por fim, aos meus colegas de curso, pelo apoio e pela colaboração ao longo desses anos de estudo.

Resumo

O entendimento dos efeitos do El Niño-Oscilação Sul (ENOS) é crucial para a agricultura brasileira, dada sua influência sobre os riscos de produção. Este trabalho teve como objetivo avaliar a importância individual das variáveis climáticas e dos municípios no desempenho do modelo, estabelecendo evidências do efeito do fenômeno sobre a classe de alto rendimento, com base nos valores SHAP. Foram analisados ciclos fixos de produção agrícola, compreendendo o período entre plantio e colheita, com duração de dez meses no estado do Rio Grande do Sul e oito meses em Goiás, nas culturas de arroz, feijão e soja. Os ciclos referentes a 1961–2019 foram agrupados nas fases do El Niño, La Niña e Neutro, empregando o método k-means funcional. O estudo concentrou-se no período de 1990 a 2019, totalizando 29 ciclos agrícolas, e testou a hipótese de que as três fases do fenômeno produzem efeitos semelhantes, por meio da Análise de Variância Funcional (FANOVA) com nível de significância de 5%. Essa análise considerou as variáveis de temperatura máxima, mínima e precipitação acumulada semanal. Os resultados mostraram evidências significativas de influência climática apenas no Rio Grande do Sul, o que motivou a realização de testes múltiplos para identificar quais fases diferem entre si. Também foi avaliado o impacto do fenômeno na produtividade das culturas por meio do algoritmo Random Forest, utilizando a classificação do rendimento categorizado por k-means clássico. Os principais resultados indicam que, no Rio Grande do Sul, a influência do El Niño-Oscilação Sul sobre as variáveis climáticas é marcante, refletindo em diferenças entre as fases analisadas. Verificou-se, pelos valores SHAP, que os fatores municipais predominam no desempenho do modelo, enquanto as variáveis climáticas apresentam menor importância, porém relevante, contribuição. Conclui-se que o fenômeno exerce impactos climáticos e produtivos significativos no Rio Grande do Sul e que, apesar de sua complexidade, os aspectos regionais municipais desempenham papel fundamental na previsão do rendimento agrícola. Esses resultados contribuem para o desenvolvimento de estratégias adaptativas frente à variabilidade climática associada ao El Niño-Oscilação Sul.

Palavras-chave: El Niño-Oscilação Sul (ENOS), Análise de Dados Funcionais, Ciclo de produção agrícola, Random Forest, valores SHAP.

Abstract

Understanding the effects of the El Niño-Southern Oscillation (ENSO) is crucial for Brazilian agriculture, given its influence on production risks. This study aimed to evaluate the individual importance of climatic variables and municipalities in model performance, establishing evidence of the phenomenon's effect on the high-yield class based on SHAP values. Fixed agricultural production cycles were analyzed, covering the period between planting and harvesting, lasting ten months in the state of Rio Grande do Sul and eight months in Goiás, for rice, bean, and soybean crops. Cycles from 1961 to 2019 were grouped according to El Niño, La Niña, and Neutral phases using the functional k-means method. The study focused on the period from 1990 to 2019, totaling 29 agricultural cycles, and tested the hypothesis that the three phases of the phenomenon produce similar effects through Functional Analysis of Variance (FANOVA) at a 5% significance level. This analysis considered the variables of maximum temperature, minimum temperature, and accumulated weekly precipitation. The results showed significant evidence of climatic influence only in Rio Grande do Sul, which motivated the performance of multiple tests to identify which phases differ from each other. The impact of the phenomenon on crop productivity was also evaluated using the Random Forest algorithm, applying yield classification categorized by classical k-means. The main results indicate that in Rio Grande do Sul, the influence of the El Niño-Southern Oscillation on climatic variables is marked, reflecting differences among the analyzed phases. SHAP values showed that municipal factors predominate in model performance, while climatic variables have lesser but still relevant contribution. It is concluded that the phenomenon exerts significant climatic and productive impacts in Rio Grande do Sul and that, despite its complexity, municipal regional aspects play a fundamental role in forecasting agricultural yield. These results contribute to the development of adaptive strategies in response to climate variability associated with the El Niño-Southern Oscillation.

Keywords: El Niño-Southern Oscillation (ENSO), Functional Data Analysis, Agricultural production cycle, Random Forest, SHAP values.

Lista de figuras

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

ENOS	El Niño–Oscilação Sul.
ADF	Análise de Dados Funcionais.
FANOVA	Análise de Variância Funcional.

Sumário

Introdução

Nos dias atuais, o El Niño–Oscilação Sul (ENOS) é conhecido como um fenômeno climático natural e complexo, impulsionado pela interação entre o oceano e a atmosfera no Pacífico Equatorial, caracterizado pelas fases El Niño, La Niña e Neutro conforme Song et al. (2019). Seu estudo é essencial em diversas áreas, como agricultura, climatologia, saúde pública e oceanografia, devido aos seus impactos severos e de ampla abrangência, especialmente na América do Sul, onde se encontra o Brasil (Anderson et al., 2018). Esses impactos podem incluir inundações urbanas, destruição de plantações e intensificação de secas, dificultando a previsão de sua variabilidade e dos efeitos sobre o clima (Cai et al., 2020).

Segundo Zhou, Liu e Cheng (2020), a anomalia do ENOS pode provocar efeitos na variabilidade climática, afetando variáveis como temperatura e precipitação. Isso desperta padrões característicos para cada fase do ENOS, podendo intensificar seus valores para acima ou abaixo do padrão geralmente observado em condições climáticas neutras. Recentemente, no Estado do Rio Grande do Sul, no ano de 2024, um terrível desastre provocado pelo grande volume de precipitação associado à fase El Niño do ENOS causou mortes e destruição, surpreendendo a população, inclusive agricultores, gerando perdas inimagináveis ao país (Morengo et al., 2024).

Dessa forma, o foco do estudo se concentrou nos estados de Goiás e Rio Grande do Sul, o que juntos representa aproximadamente 27% da produção nacional de arroz, feijão e soja, conforme os dados apresentados no painel do IBGE (2024). Assim, estabelecer a janela em que ocorrem as atividades agrícolas dessas culturas em cada Estado é um ponto de partida importante, sendo delimitada para o Estado do Rio Grande do Sul entre o início de setembro e o final de junho do ano seguinte, com duração de 10 meses. Para o Estado de Goiás entre o início de outubro e final de maio, com duração de 8 meses conforme a (Conab, 2022).

O Rio Grande do Sul lidera a produção nacional de arroz irrigado, com 8,3 milhões de toneladas, equivalendo a 74,3% do total produzido no país, além de ocupar a quarta posição na produção de soja, com 14,3 milhões de toneladas (8,4% do total nacional). A produção de feijão-comum no estado, especialmente do tipo comercial “preto”, também é significativa, totalizando 56,2 mil toneladas, ou 7,1% da produção brasileira (Conab, 2025). Já o estado de Goiás destaca-se como o terceiro maior produtor de grãos do Brasil, com 35 milhões de toneladas na safra 2024/25 (cerca de 10% da produção nacional), onde a soja predomina com 58% do volume, seguida por milho (36%), sorgo (4%) e feijão-comum (0,9%) (Conab, 2025).

Assim, a análise de dados funcionais é proposta como uma alternativa robusta para compreender a complexidade e a dinâmica dos fenômenos associados ao ENOS, sendo o k-means funcional o método inicial deste trabalho. Esse procedimento tem o papel de preservar a linha temporal das observações, sem a necessidade de uma redução excessiva dos dados, o que, neste

caso, poderia amplificar o ruído (Ullah e Finch, 2013). Essa abordagem permite identificar e avaliar a homogeneidade das fases do ENOS ao longo do tempo, em cada ciclo agrícola, verificando de forma condicional seus efeitos sobre o rendimento das três culturas por meio de um modelo baseado em árvores. O foco deste estudo está na interpretabilidade proporcionada pelos valores SHAP, por meio dos quais se busca avaliar a importância individual das variáveis climáticas e dos municípios no desempenho do modelo, estabelecendo evidências do efeito do fenômeno sobre a classe de alto rendimento. Dessa forma, esta pesquisa fornece uma compreensão detalhada dos efeitos do ENOS sobre temperatura e precipitação no rendimento agrícola, contribuindo para identificar padrões de importância associados às condições de alto rendimento sob a influência do ENOS.

1 Revisão Bibliográfica

O presente capítulo dedica-se à revisão bibliográfica, organizada de acordo com o desenvolvimento do estudo, com a apresentação do tema; abordagens estatísticas para dados funcionais; critérios para prevenção do erro tipo I em testes múltiplos, modelo de regressão local, agrupamento k-means clássico e, por fim, modelo Random Forest.

1.1 Fenômeno El Niño-Oscilação Sul (ENOS)

O fenômeno El Niño–Oscilação Sul, tema central deste trabalho, foi inicialmente identificado por Walker (1925), que mostrou correlação nas variações sazonais das condições atmosféricas, levando à definição da Oscilação Sul como uma flutuação na pressão atmosférica em escala global, caracterizada por um vaivém entre o Oceano Pacífico e a região do Oceano Índico/Indonésia. No entanto, o estudo de Walker era essencialmente descritivo e não explicava os mecanismos físicos subjacentes.

Décadas depois, com o estudo realizado por Bjerknes (1969), unificaram o conceito, ao demonstrar que a Oscilação Sul e o fenômeno oceânico El Niño são partes de um sistema acoplado, mostrando que um enfraquecimento anômalo dos ventos alísios no Hemisfério Sul suprime a ressurgência equatorial, causando o aquecimento das águas no Pacífico central e oriental (El Niño). Por sua vez, esse aquecimento oceânico altera a liberação de calor para a atmosfera, intensificando as anomalias na circulação atmosférica. Foi essa descoberta do mecanismo de realimentação entre o oceano e a atmosfera que fundamentou o conceito moderno do fenômeno ENOS.

1.2 Ciclo de Produção Agrícola

Pereira, Angelocci e Sentelhas (2007) destacam que o ciclo produtivo agrícola, composto pelas etapas de plantio, desenvolvimento da cultura e colheita, é diretamente condicionado por fatores climáticos e meteorológicos. Isso sugere a necessidade de planejamento já na fase de plantio, em que a escolha da época de semeadura deve considerar o regime de chuvas e a temperatura. Durante o desenvolvimento, elementos como a radiação solar, o balanço hídrico e a umidade influenciam o crescimento e a sanidade da cultura. Por fim, a colheita deve ser planejada com base no acúmulo térmico e nas condições do tempo, de modo a evitar perdas qualitativas e quantitativas. Mostra-se, em Iizumi et al., (2014), que o ENOS é um fator climático de grande importância, uma vez que frequentemente influencia a temperatura e a precipitação sazonais, afetando assim os rendimentos agrícolas em diversos países, inclusive o Brasil.

1.3 Análise de Dados Funcionais (ADF)

Segundo Ramsay e Dalzell (1991), a Análise de Dados Funcionais é desenvolvida a partir de dados representados por funções contínuas, denotadas por $x_i(t)$, para $i = 1, \ldots, n$, definidas em um domínio $t \in T$, no qual T representa um intervalo real. Assim, cada x_i corresponde a um ponto em um espaço de funções, também denominado espaço de Hilbert (H). De acordo com a definição apresentada por Kreyszig (1991), o espaço de Hilbert é descrito na literatura como um espaço vetorial completo com produto interno. Essa estrutura amplia as possibilidades de aplicação da ADF, inclusive em espaços de dimensão finita.

A ADF tem sido amplamente aplicada em diversos contextos. Um estudo recente conduzido por Yildirim, Franco-Pereira e Lillo (2025), envolvendo um problema com múltiplas variáveis, apresentou uma metodologia de Análise de Componentes Principais Funcionais Multivariada (MFPCA) para dados funcionais multivariados, aplicada ao monitoramento de condição e à classificação de múltiplas falhas em sistemas hidráulicos. Os escores obtidos, utilizados como variáveis independentes, resultaram em uma precisão de classificação superior do modelo Random Forest em comparação com as seis abordagens sem ADF avaliadas no estudo, evidenciando o ganho que essa abordagem pode oferecer.

Vale notar que uma das principais vantagens da Análise de Dados Funcionais reside em sua flexibilidade de modelagem. Em vez de se restringir a formas paramétricas pré-definidas, que limitam a relação entre variáveis a um número fixo de parâmetros, a ADF adota uma abordagem não paramétrica. Nesse contexto, as funções são representadas por combinações de bases flexíveis, como, por exemplo, bases de splines cúbicos, nas quais a complexidade do modelo é controlada pelos próprios dados, geralmente por meio de penalização (Ferraty e Vieu, 2006). Dessa forma, a função estimada configura-se como uma alternativa viável quando modelos paramétricos convencionais não conseguem representar adequadamente tais padrões.

1.3.1 Modelos de Agrupamento: K-means Clássico e Funcional

1.3.1.1 K-means Clássico

O algoritmo K-means de Hartigan e Wong (1979) é um método iterativo desenvolvido para particionar M pontos em K grupos, com o objetivo de minimizar a soma dos quadrados dentro de cada cluster, produzindo agrupamentos internamente homogêneos. A análise de Lloyd (1982) consolidou os fundamentos teóricos do K-means e destacou seus principais desafios, tornando-se referência central na literatura. O autor demonstrou que a função de custo é não convexa e pode apresentar múltiplos mínimos locais, dependendo da distribuição de probabilidade subjacente aos dados. Seu "Método I", de natureza iterativa, alterna entre as etapas de atribuição dos pontos e atualização dos centróides, mas não garante convergência ao ótimo global. Essa limitação motiva a prática amplamente adotada de executar o algoritmo diversas vezes com diferentes

inicializações aleatórias.

A contribuição de Hartigan e Wong (1979) para a eficiência computacional do método k-means estabeleceu padrões de robustez e usabilidade que influenciaram gerações de implementações. O algoritmo inclui diagnósticos de falha, como a detecção de clusters vazios e a verificação de limites para o número de clusters, assegurando maior confiabilidade na prática. Testes comparativos em conjuntos de dados simulados demonstraram que a abordagem proposta atinge soluções localmente ótimas, com garantias que não eram oferecidas por algoritmos anteriores, e sugeriram a inicialização baseada na distância ordenada à média global para mitigar problemas decorrentes de inicializações aleatórias.

Atualmente, o método de agrupamento k-means desempenha um papel relevante em diversas aplicações. Um exemplo é seu uso no processo de segmentação de regiões de interesse (ROI) em folhas de plantas, com o objetivo de isolar áreas saudáveis das áreas doentes (Yashodha et al., 2025). Outra implementação, apresentada por Costa-Neto et al. (2024), envolve a identificação de clusters ambientais com base em dados históricos, representando regiões com condições ecológicas semelhantes e, consequentemente, com potenciais produtivos próximos. Essa aplicação auxilia na estratificação das regiões de produção em zonas de melhoramento mais homogêneas, reforçando o potencial do método em soluções voltadas às ciências agrárias.

1.3.1.2 K-means Funcional

O Algoritmo k-Means para Funções é realizado da seguinte forma. Suponha que os dados consistam em n funções, $\mathbf{f}_1(t), \ldots, \mathbf{f}_n(t)$, definidas em um intervalo $[T_1, T_2]$, onde T_1 e T_2 representam, respectivamente, o início e o fim de um intervalo contínuo no qual as funções existem. A associação de grupo para a i-ésima observação, $f_i \in \{1, 2,...,k\}$, é um dado desconhecido que será identificado após a convergência (Tarpey e Kinateder, 2003). O método k-means funcional pode ser considerado uma adaptação do k-means clássico, conforme descrito por Hartigan e Wong (1979), cuja formulação teórica foi desenvolvida para lidar com funções suaves e estimáveis no espaço de Hilbert, quando aplicado a uma matriz de valores discretizados.

Este feito, para o método k-means funcional, é viabilizado pela conversão de matrizes de dados discretizados em objetos funcionais. Esse processo foi implementado computacionalmente no pacote fda.usc (Febrero; Bande; Fuente, 2012), considerado uma das principais bibliotecas para ADF, cuja classe “fdata” foi especificamente desenvolvida para o tratamento de dados funcionais discretizados.

O estudo de Orozco, Ortiz e Ospina-Tascón (2025) destaca a relevância do algoritmo K-means funcional no contexto médico, evidenciando-o como o método de agrupamento funcional mais utilizado em sua revisão sistemática. A análise identificou que o K-means foi aplicado em 12 dos 93 artigos revisados, correspondendo a aproximadamente 12,9% do total de estudos publicados entre 1995 e 2024 que utilizaram Análise de Dados Funcionais. Este destaque reforça a utilidade do método para identificar subgrupos de pacientes com base em padrões temporais

de variáveis clínicas, como trajetórias de glicose, frequência cardíaca e sinais vitais, permitindo estratificações clínicas significativas e potencialmente orientando intervenções personalizadas.

1.3.2 Análise de Variâcia Funcional (FANOVA)

A FANOVA é uma análise de variância para dados funcionais, utilizada para testar a hipótese nula de que as funções médias $m_k(t)$ dos múltiplos grupos de dados funcionais são idênticas, conforme descrito por Górecki e Smaga (2015). A hipótese a ser testada é:

$$H_0 : m_1(t) = \cdots = m_k(t), \quad t \in T, \tag{1.1}$$

sendo a alternativa a negação de H_0.

Essa abordagem é inerentemente não paramétrica, o que significa que o teste de permutação baseado em uma representação de função base (FP) não exige suposições rígidas sobre a distribuição de probabilidade subjacente dos dados funcionais, diferentemente da ANOVA tradicional aplicada a dados não funcionais.

O procedimento de permutação consiste em gerar um grande número de permutações aleatórias dos dados observados. Para cada permutação, calcula-se a estatística F, e o valor de p corresponde à proporção de estatísticas F obtidas das permutações que são maiores ou iguais à estatística F observada nos dados originais.

O valor de K, que representa o número de funções de base utilizadas na representação da função, pode ser selecionado individualmente para cada observação com base em um critério de informação. O Critério de Informação Bayesiano (BIC) é uma das opções disponíveis para determinar o valor ótimo de K. Além disso, ao empregar bases como b-splines, o valor máximo de K é restrito a ser menor ou igual ao número de pontos de discretização (Górecki; Smaga, 2015).

1.4 Comparação Múltipla: Método de Holm-Bonferroni

Segundo Abdi (2010), quanto mais testes estatísticos são realizados, maior é a probabilidade de rejeitar a hipótese nula quando ela é verdadeira, ou seja, de ocorrer um falso alarme, também conhecido como erro tipo I. Esse problema é descrito como inflação do nível alfa. O autor complementa que, para uma família de C testes, a probabilidade de cometer ao menos um erro tipo I é dada por $1 - (1 - \alpha)^C$, resultando, por exemplo, em uma probabilidade de aproximadamente 0,226 quando $\alpha = 0{,}05$ e $C = 5$. Esse valor demonstra que a probabilidade real de rejeitar a hipótese nula quando ela é verdadeira em pelo menos um dos testes é superior ao nível de $\alpha = 0{,}05$ definido para testes individuais.

Thompson (1998) argumenta que, se a ausência de correção para múltiplas comparações fosse aceita como prática comum, ou se os pesquisadores pudessem explorar livremente os dados antes de definir suas hipóteses, os valores de p publicados perderiam ainda mais credibilidade. Tal procedimento equivaleria, segundo o autor, a permitir a publicação de coincidências apresentadas de forma enganosamente científica.

Dessa forma, o método proposto por Holm (1979) consiste em um procedimento sequencial de comparações múltiplas que ordena os valores de p e ajusta os níveis de significância, configurando-se como uma alternativa menos conservadora ao método clássico de Bonferroni para o controle da taxa de erro tipo I em testes de hipóteses múltiplas.

1.5 Regressão Local (*LOESS*)

Conforme Gubels e Prosdocimi (2010), a regressão local utiliza um método de suavização não paramétrico para modelar a relação entre as variáveis explicativas e a variável resposta, também empregando o método dos mínimos quadrados, como ocorre no modelo de regressão linear paramétrico. No entanto, diferentemente da regressão linear, a *LOESS* não realiza um ajuste global, mas sim local. Sua principal característica está na flexibilidade de empregar funções polinomiais ajustadas localmente, considerando apenas as observações vizinhas a cada ponto, o que permite capturar variações mais sutis nos dados, especialmente em situações em que o comportamento não é linear.

Dessa forma, o processo de determinação da vizinhança dos dados que influenciará o valor ajustado em um ponto-alvo x_k é obtido a partir de um ajuste polinomial de grau d aos dados, utilizando o método dos mínimos quadrados ponderados. A ponderação atribuída a cada ponto de dado, situado em (x_i, y_i), é definida de modo que seja maior quando x_i estiver próximo de x_k e menor quando estiver mais distante (Cleveland, 1979).

Uma aplicação da regressão local foi apresentada por Simpson e Haggard (2018) na avaliação de tendências em cursos d'água para a detecção de poluentes ajustada por vazão. O uso desse modelo não paramétrico foi essencial para isolar a influência de variáveis hidrológicas nas concentrações de constituintes, com automatização baseada em validação cruzada, o que permitiu a detecção de tendências monotônicas reais na qualidade da água, isolando o efeito de intervenções de gestão, como práticas de conservação da variabilidade hidrológica natural. Há uma perspectiva semelhante quanto ao uso do modelo de regressão local neste trabalho, porém com aplicação voltada à estimativa das curvas de tendência tecnológica em dados de rendimento.

Segundo o Painel da Embrapa (2021), observam-se diferenças tanto na produção quanto na área plantada ao longo dos anos, entre 1974 e 2021. Ramalho, Marques e Lemos (2021) realizaram uma revisão histórica e analítica do melhoramento genético de plantas no Brasil nas últimas cinco décadas, destacando marcos científicos e institucionais, como a criação da Embrapa, em 1973, e as tecnologias que transformaram o país em uma das maiores potências

agrícolas do mundo. Esse cenário apresenta mudanças expressivas ao longo do tempo, o que torna necessário estudá-lo sob condições climáticas estáveis, em um contexto isento da influência da evolução tecnológica, mantendo apenas as variações climáticas nos dados de rendimento das culturas. É nesse contexto que se insere o modelo de regressão local.

1.6 Random Forest (RF)

O modelo Random Forest, segundo James et al. (2013), é uma técnica de aprendizado supervisionado baseada em árvores de decisão, aplicada a problemas de classificação e regressão. O modelo utiliza a abordagem de Bagging, que combina um grande número de árvores de decisão individuais para gerar uma única resposta.

Essa robustez do Random Forest, especialmente via Bagging, revela-se particularmente valiosa em cenários reais com violações de pressupostos paramétricos, como demonstrado por Chowdhury et al. (2022). Nessa aplicação, os autores compararam modelos de regressão linear múltipla e Random Forest em um problema de regressão com dados assimétricos e multicolineares de desempenho de baterias. O modelo Random Forest, por ser não paramétrico, não foi afetado pelas suposições de normalidade e multicolinearidade. Além disso, o método Bagging, utilizado no RF, foi essencial para reduzir o overfitting, mostrando-se uma opção superior em conjuntos de dados em que as suposições da regressão linear múltipla são violadas.

É importante destacar a necessidade de construir uma grade de hiperparâmetros para a otimização do modelo. Esse processo é realizado por meio de tunagem, com o objetivo de melhorar a acurácia do modelo e permitir controle computacional adequado. Os hiperparâmetros são definidos antes do início do treinamento, sendo estabelecidos com base nas características dos dados e na capacidade do algoritmo de aprendizado (Agrawal, 2021).

De acordo com Agrawal (2021) e Hastie et al. (2009), os principais hiperparâmetros são:

- **num.trees**: número de árvores combinadas no modelo.
- **mtry**: número de variáveis explicativas aleatoriamente selecionadas por divisão.
- **min.node.size**: parâmetro que controla o número mínimo de observações em um nó terminal (folha).
- **max.depth**: profundidade máxima das árvores de decisão que compõem o modelo.

Para avaliar a importância das variáveis explicativas (*features*), existem diversos métodos; contudo, neste trabalho foram utilizados os valores Shapley (ou valores SHAP). Segundo Strumbelj e Kononenko (2014), os valores SHAP constituem um método de explicação de modelos de previsão baseado no conceito de valor de Shapley da teoria dos jogos cooperativos. O método atua perturbando sistematicamente todos os subconjuntos possíveis da variável explicativa de

entrada e observando a variação na previsão do modelo, atribuindo a cada *feature* uma contribuição marginal média decorrente de sua presença em todas as coalizões possíveis. A principal vantagem desse método é sua capacidade de lidar com modelos não aditivos e complexos, nos quais métodos que perturbam uma única feature por vez falham. Além de explicar previsões individuais, o método pode ser estendido para mensurar a importância global das *features*, por meio do cálculo do desvio padrão de suas contribuições locais, identificando quais são mais influentes (Strumbelj e Kononenko, 2014).

2 Metodologia

2.1 Origem dos Dados

De início, estabeleceu-se um intervalo substancial de dez meses, entre setembro e julho, para representar o ciclo de produção agrícola das culturas de arroz, feijão e soja no Estado do Rio Grande do Sul. No caso do Estado de Goiás, o intervalo corresponde a oito meses, abrangendo o período de outubro a maio para as mesmas três culturas, conforme informações da Companhia Nacional de Abastecimento (CONAB, 2022).

Para as análises, foram utilizados três conjuntos de dados. O primeiro, usado no método K-means funcional, com dados fornecendo informações sobre a temperatura da superfície do Oceano Pacífico para o período de 1961 a 2019. Esses dados mensais foram obtidos no Climate Prediction Center (CPC), órgão da National Oceanic and Atmospheric Administration (NOAA, 2020), que define o Oceanic Niño Index (ONI) como a média das anomalias trimestrais da TSM na região Niño, sendo utilizado para identificar os eventos de El Niño e La Niña. O El Niño é caracterizado pelo aquecimento anômalo da TSM, superior a +0.5 °C, enquanto o La Niña refere-se ao resfriamento anômalo da TSM, superior a -0.5 °C. Já os períodos neutros são aqueles em que há ausência de eventos significativos de El Niño ou La Niña. Esses dados das anormalias trimestrais, estão disponíveis em: <https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php>.

O segundo conjunto de dados apresenta variáveis meteorológicas em unidades diárias de temperatura máxima (C°), temperatura mínima (C°) e precipitação (mm) de municípios para os Estados de Goiás e Rio Grande do Sul. Com a utilização de dados entre os períodos de (1990-2019), obtidos junto ao Instituto Nacional de Meteorologia (INMET, 2019).

Por fim, utilizaram-se dados anuais de rendimento agrícola de arroz, feijão e soja, em quilogramas por hectare ($kg\,ha^{-1}$), referentes aos municípios dos estados de Goiás e Rio Grande do Sul mencionados anteriormente. Esses dados foram extraídos diretamente do Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2024), abrangendo o período de 1991 a 2019 e considerando o ano de consolidação dos rendimentos das culturas após o ciclo de produção agrícola, que constitui o objeto de avaliação deste estudo.

2.2 Critério de Tratamentos de Dados e Ajuste de Escala

As descrições contidas na documentação que acompanhava os conjuntos de dados meteorológicos informavam a porcentagem de observações imputadas em cada base de dados, por município e por Estado. A decisão adotada foi filtrar estações meteorológicas que apresentassem

uma porcentagem máxima de 30% de tolerância para observações estimadas, visando capturar mais estações e, ao mesmo tempo, manter uma qualidade razoável de observações íntegras coletadas efetivamente. Nessa condição, no Estado de Goiás, foram selecionados 7 municípios com estações de dados de temperatura e 2 com dados de precipitação. No Rio Grande do Sul, foram considerados 16 municípios para temperatura e 46 para precipitação. Para mais informações, dados complementares e materiais adicionais, acesse o repositório pessoal disponível em: <https://github.com/UssMV/Materiais-Monografia-TCC-Marcos-#>. Com as localidades das estações meteorológicas pertencentes aos Estados:

Goiás

- **Temperatura:** Aragarcas, Goiás, Formosa, Goiânia, Catalão.
- **Precipitação:** Inhumas, Goiânia.

Rio Grande do Sul

- **Temperatura:** Iraí, São Luiz Gonzaga, Cruz Alta, Passo Fundo, Bom Jesus, Uruguaiana, Santa Maria, Bento Gonçalves, Caxias do Sul, Torres, Triunfo, Encruzilhada do Sul, Porto Alegre, Bagé, Pelotas, Rio Grande.
- **Precipitação:** Três de Maio, Canguçu, Porto Lucena, Porto Tarumã, Passo São José, Dom Pedrito, Paraíso, Erebanho, Passo Tainhas, Coqueiros do Sul, Ernesto Alves, Passo do Sarmento, Cacequi, Serra do Pinto, Encantado, Dona Francisca, Nova Palmira, Giruá, Garruchos, Prata, São Lourenço do Sul, Passo Viola, Carazinho, Cachoeira Santa Cecília, Jaguari, Rosário do Sul, Santa Clara do Ingaí, Lavras do Sul, Alto Uruguai, Passo do Prata, Santo Augusto, Antônio Prado, Sananduva, Passo Faxinal, Passo da Guarda CEEE, Linha Céson, Ilópolis, Passo Major Zeferino, Plano Alto, Passo Migliavaca, Seca, Passo do Mendonça, Vila Três Passos, Conceição, Paim Filho.

No caso dos dados anuais de rendimento, adotou-se como critério a seleção de municípios que apresentassem registros de produtividade das culturas com plantio e desenvolvimento nos anos de interesse, de 1991 a 2019, abrangendo todas as culturas analisadas. Assim, foram selecionados 68 municípios para os dados de arroz, 165 para soja e 182 para feijão.

Para cada município, período e ciclo agrícola, os dados meteorológicos diários foram agregados em semanas de 7 dias. A agregação dos dados diários em semanais foi realizada da seguinte forma: para as temperaturas, consideraram-se os valores máximo e mínimo de cada semana; para a precipitação, somaram-se os valores diários. Para o processamento e análise dos dados, foi utilizado o software R (R Core Team, 2025).

2.3 K-Means Funcional

A aplicação do K-Means Funcional considerou três clusters, com o objetivo de identificar as três fases (El Niño, La Niña e Neutro), utilizando um total de 100 inicializações para obter um melhor agrupamento e evitar o viés de escolha inicial do centróide. O método de clusterização é definido formalmente no espaço de funções quadrado-integráveis, $L^2(T)$, pertencente ao espaço de Hilbert. Nesse espaço de funções $L^2(T)$, um conjunto de médias de cluster determina uma partição do espaço de acordo com a distância mínima (Tarpey e Kinateder, 2003).

O procedimento iterativo consiste nas quatro etapas a seguir:

1. **Inicialização:** Para agrupar as funções usando o algoritmo k-means, inicia-se com k médias de cluster iniciais, arbitrárias e distintas: $m_1(t), \ldots, m_k(t)$.

2. **Passo de Atribuição de Cluster:** Cada função $\mathbf{f}_i(t)$ é atribuída ao cluster cuja média ($\mathbf{m}_j$) esteja mais próxima, de acordo com a distância $L^2[T_1, T_2]$ ao quadrado:

$$L^2[T_1, T_2] = \min_j \left[\int_{T_1}^{T_2} (m_j(t) - f_i(t))^2 dt \right].$$

3. **Atualização do Centróide:** As médias de cluster são atualizadas calculando-se a média aritmética de todas as funções pertencentes ao cluster correspondente, sendo $\mathbf{n}_j$ o número de funções no j-ésimo cluster:

$$m_j(t) = \frac{1}{\mathbf{n}_j} \sum_{\mathbf{f}_i \in j} \mathbf{f}_i(t).$$

   Após todas as funções $\mathbf{f}_i(t)$ serem atribuídas a um cluster, as médias são atualizadas.

4. **Convergência:** O procedimento prossegue até que nenhuma função mude de cluster, finalizando o algoritmo.

2.4 Algoritmo da Análise de Variância Funcional (FANOVA)

Em seguida, utiliza-se a FANOVA para avaliar se as funções presentes entre os grupos El Niño, La Niña e Neutro são estatisticamente iguais (1.1), quando avaliadas nas variáveis climáticas. Conforme Górecki e Smaga (2015), as funções contínuas $X_{ij}(t)$ são representadas por uma soma ponderada de um número finito K de funções de base ortonormais $\phi_l(t)$:

$$X_{ij}(t) = \sum_{l=0}^{K} c_{ijl}\phi_l(t); \quad t \in T,$$

onde c_{ijl} são os coeficientes aleatórios. Para o teste FP, o sistema B-spline é escolhido como o sistema de funções de base ortonormais. Sendo que i é o índice que denota o grupo, variando de

1 ao k (número total de grupos), e j é o índice que denota a função no grupo específico.

As funções médias amostrais de grupo $\overline{X_i}(t)$ em (2.1) e a função média geral da amostra $\overline{X}(t)$ vista em (2.2) podem ser escritas em notação matricial em termos dos vetores de coeficientes $c_{ij} = (c_{ij0}, c_{ij1}, \ldots, c_{ijK})'$:

$$\overline{X_i}(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} c'_{ij}\phi(t), \tag{2.1}$$

$$\overline{X}(t) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} c'_{ij}\phi(t), \tag{2.2}$$

onde $\phi(t) = (\phi_0(t), \phi_1(t), \ldots, \phi_K(t))'$, n representa o número total de funções em todos os grupos combinados e l é o índice que denota a função de base.

Em seguida, é realizado o cálculo da Estatística de teste (2.3), para cada variável climática na FANOVA para testar se há diferença significativa entre os três grupos. Com o valor-p, obtido através de métodos de permutação, que reembaralham aleatoriamente as funções entre os grupos para simular a distribuição da estatística sob a hipótese nula.

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i ||\overline{X}_i - \overline{X}||_2^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} ||X_{ij} - \overline{X}_i||_2^2}, \tag{2.3}$$

onde $||f||_2^2 = \int_T f^2(t)dt$ para $f \in L_2(T)$.

Quando diferenças significativas são encontradas nas estatísticas de teste, surge a necessidade de verificar qual ou quais grupos (El Niño, La Niña e Neutro) da variável climática se diferenciam. Isso leva à aplicação de testes de comparação múltipla. Nessa etapa, o procedimento é realizado comparando todas as combinações possíveis entre os grupos, duas a duas, por meio do método de Holm-Bonferroni, conforme (Holm, 1979):

Sejam $Y_i = Pr_{H_i}(T_i \geq t_i)$ os valores p associados a cada hipótese H_i, $i = 1, \ldots, n$. Ordenam-se os valores p de forma crescente:

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}.$$

A hipótese correspondente ao menor valor-p, $H_{(1)}$, é rejeitada se

$$Y_{(1)} \leq \frac{\alpha}{n},$$

Caso a hipótese anterior seja rejeitada, passa-se à próxima: $H_{(2)}$ é rejeitada se

$$Y_{(2)} \leq \frac{\alpha}{n-1},$$

e assim sucessivamente, de modo que, no passo j, $H_{(j)}$ é rejeitada se

$$Y_{(j)} \leq \frac{\alpha}{n-j+1}. \tag{2.4}$$

Esta abordagem apresenta maior poder estatístico que o método de Bonferroni clássico, mantém o controle do erro tipo I no nível α, além de ser amplamente aplicável a modelos paramétricos e não paramétricos, conforme demonstrado por (Holm, 1979).

2.5 Modelo de Regressão Local (*LOESS*)

A abordagem do modelo de regressão local foi utilizada para estimar a curva de evolução tecnológica ao longo dos anos, com o objetivo de remover essa tendência dos dados de rendimento das culturas. Segundo Gubels e Prosdocimi (2010), essa metodologia considera uma função de regressão desconhecida $m(x)$. A técnica assume que essa função pode ser aproximada por uma função polinomial de grau p em uma vizinhança de um ponto fixo z. Isso é feito por meio de uma expansão de Taylor até a ordem p:

$$m(x) \approx m(z) + m'(z)(x-z) + \cdots + \frac{m^{(p)}(z)}{p!}(x-z)^p \equiv \beta_0 + \beta_1(x-z) + \cdots + \beta_p(x-z)^p,$$

em que $\beta_j = \frac{1}{j!}m^{(j)}(z)$, para $j = 0, \ldots, p$.

Para estimar os parâmetros $\beta_0, \ldots, \beta_p$ em um dado ponto z, resolve-se o problema de minimização dos mínimos quadrados ponderados:

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - z)^j \right)^2 K\left(\frac{X_i - z}{h} \right),$$

cujos argumentos são descritos a seguir:

- Y_i: valor observado da variável resposta para a i-ésima observação;
- X_i: valor da covariável para a i-ésima observação;
- z: ponto fixo em torno do qual a regressão local é realizada;
- β_j: coeficientes do polinômio local a serem estimados;
- p: grau do polinômio local usado na aproximação;
- $K(\cdot)$: função kernel, que atribui pesos às observações com base na proximidade de X_i a z (valores próximos de z recebem maior peso);

- h: parâmetro de largura de banda, que controla o tamanho da vizinhança em torno de z (um h maior inclui mais observações no ajuste local).

Em notação matricial, o problema de minimização pode ser reescrito como $\min_\beta (Y - X_D\beta)^T W (Y - X_D\beta)$, em que Y é o vetor coluna de observações, β é o vetor coluna de parâmetros, X_D é a matriz de delineamento e W é a matriz diagonal de pesos. A solução para o problema de minimização é dada por:

$$\hat{\beta} = (X_D^T W X_D)^{-1} X_D^T W Y,$$

desde que exista a inversa da matriz $X_D^T W X_D$. A estimativa da função de regressão no ponto z é $\hat{m}(z) = \hat{\beta}_0$. O procedimento de ajuste da produtividade, com base em Heinemann e Sentelhas (2011), pode ser descrito nas etapas a seguir:

Etapa 1: Cálculo do desvio relativo

O desvio relativo (RD_i^n) é calculado para cada município e ano, representando o afastamento da produtividade observada em relação à tendência tecnológica:

$$RD_i^n = \frac{x - \bar{y}}{\bar{y}}, \tag{2.5}$$

em que:

- RD_i^n: desvio relativo do ano i em relação ao ano n = 2019;
- x: produtividade observada (kg ha^{-1});
- $\bar{y}$: produtividade predita pela regressão local (kg ha^{-1}).

Etapa 2: Ajuste do rendimento para o último ano de referência em cada período, utilizando o desvio relativo:

$$AY_1^n = (RD_1^n + 1) \cdot \bar{y}_n, \tag{2.6}$$

em que:

- AY: produtividade ajustada (kg ha^{-1});
- $\bar{y}_n$: produtividade predita por *LOESS* no ano n = 2019.

2.6 Algoritmo de Otimização do K-means Clássico

O algoritmo K-Means foi utilizado para classificar os rendimentos de cada cultura em dois grupos distintos: alto e baixo desempenho produtivo, os quais foram posteriormente utilizados no modelo de classificação Random Forest. Esse método iterativo busca soluções ótimas locais para formar os dois grupos. Conforme a definição de Pena, Lozano e Larrañaga (1999), a função de critério, denotada por F, é o que o algoritmo K-Means tenta minimizar. Ela é definida como a soma dos quadrados das distâncias euclidianas L_2 entre cada ponto dado e o centro (centroide) do cluster ao qual ele pertence.

Matematicamente, a função de critério F para uma partição de um banco de dados em K clusters, $\{C_1, \ldots, C_K\}$, é dada por:

$$F(\{C_1, \ldots, C_K\}) = \sum_{i=1}^{K} \sum_{j=1}^{k_i} \|\mathbf{w}_{ij} - \bar{\mathbf{w}}_i\|^2,$$

onde:

- K é o número de clusters;
- k_i é o número de objetos no cluster i;
- $\mathbf{w}_{ij}$ é o j-ésimo objeto do i-ésimo cluster;
- $\bar{\mathbf{w}}_i$ é o centroide (média) do i-ésimo cluster, calculado como:

$$\bar{\mathbf{w}}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \mathbf{w}_{ij}.$$

Apesar de o algoritmo ser robusto e garantir a convergência para um mínimo local, essa solução final é sensível às condições iniciais, como o agrupamento inicial e a ordem dos dados. Isso significa que diferentes pontos de partida podem levar a diferentes soluções locais (Pena, Lozano e Larrañaga, 1999). Diante disso, o algoritmo foi executado 25 vezes com centroides iniciais distintos.

2.7 Modelo Random Forest

O modelo Random Forest consiste em um conjunto de classificadores do tipo árvore, em que cada árvore é construída a partir de vetores aleatórios independentes, extraídos da mesma distribuição para todas as árvores da floresta. Essas árvores são geradas de forma independente, com suas decisões tomadas em paralelo (Breiman, 2001). Diante disso, compreender a estrutura de uma única árvore de classificação utilizada neste trabalho é fundamental, pois ela representa

o componente básico pelo qual toda a floresta é construída e determina, em grande parte, a capacidade preditiva do modelo.

Antes da execução do modelo, foi definida uma grade de hiperparâmetros para o processo de ajuste, etapa considerada essencial para garantir o melhor desempenho, conforme descrito por Agrawal (2021). Para o parâmetro que representa o número de árvores, "num.trees", foi estabelecido um intervalo de valores entre 5 e 150. O parâmetro referente ao número de variáveis explicativas consideradas em cada divisão foi ajustado entre 1 e o total de variáveis disponíveis no conjunto de dados. Já o parâmetro "min.node.size", variou de 5 a 30, enquanto a profundidade máxima das árvores, "max.depth", foi definida no intervalo de 5 a 50.

Uma árvore de classificação é utilizada para prever uma resposta qualitativa. O processo de construção de uma árvore envolve a segmentação recursiva do espaço do preditor, criando regiões em forma de caixa. Em cada região, a escolha da previsão de classe é a moda das observações de treinamento que caem nessa região. Para o crescimento da árvore, o algoritmo utiliza a divisão binária recursiva. Em cada etapa, ele seleciona o preditor e o ponto de corte que dividem os dados em duas regiões, de forma a maximizar a pureza dos nós resultantes (James et al., 2013).

A pureza do nó é um conceito que indica o quão homogêneo é um grupo de dados dentro de um nó específico, facilitando a classificação correta dos registros. Um nó é considerado puro quando todos os exemplos nele pertencem à mesma classe. Caso contrário, ele é classificado como impuro, apresentando uma mistura de classes diferentes. O uso do índice de Gini como métrica é preferido para avaliar a pureza da divisão, pois é mais sensível do que a simples taxa de erro na orientação da construção da árvore, resultando em nós mais homogêneos e em melhor desempenho do modelo (Sharda, Voss e Suthaharan, 2019).

O índice de Gini é uma medida da impureza (ou falta de homogeneidade) entre as K classes. Para uma região m, ele é definido como:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

em que $\hat{p}_{mk}$ é a proporção de observações de treinamento na região m que pertencem à classe k. Um valor pequeno do Gini indica alta pureza do nó (James et al., 2013).

Bagging é uma técnica desenvolvida para melhorar a precisão de métodos de predição, especialmente aqueles que são instáveis, como árvores de decisão e redes neurais. O método, proposto por Breiman (1996), baseia-se na ideia de gerar múltiplas versões de um preditor a partir de diferentes subconjuntos da base de dados, obtidos por amostragem com reposição (bootstrap), e então combinar esses preditores para formar uma predição agregada.

O processo de bagging segue três passos fundamentais. Inicialmente, na etapa de amostragem bootstrap, são geradas múltiplas amostras a partir do conjunto de dados original. Cada uma dessas amostras possui o mesmo tamanho do conjunto original, mas é criada por meio de um processo de amostragem com reposição, o que permite que algumas instâncias sejam

selecionadas mais de uma vez, enquanto outras podem ser omitidas. Em seguida, ocorre o treinamento de múltiplos modelos, em que uma árvore de decisão independente é treinada para cada uma das amostras bootstrap criadas anteriormente. Por fim, no estágio de agregação das predições, os resultados individuais de todos os modelos são combinados. Especificamente, para problemas de classificação, a predição final é determinada por meio de uma votação majoritária entre todas as árvores de decisão treinadas.

Em seguida, com o modelo ajustado, foi aplicada uma última metodologia aos valores SHAP com o objetivo de determinar a importância individual das variáveis climáticas e do município na previsão média do modelo de classificação. O método de valores SHAP por aproximação calcula a contribuição de cada variável explicativa por meio de amostragem, evitando a complexidade exponencial do cálculo exato. A formulação teórica do valor de Shapley para a contribuição da i-ésima variável, que permite a aproximação amostral, é dada por:

$$\varphi_i(x) = \frac{1}{n!} \sum_{\mathcal{O} \in \pi(N)} \sum_{w \in \mathcal{X}} p(w) \cdot \left[f(w_{[w_j = x_j, j \in Pre^i(\mathcal{O}) \cup i]}) - f(w_{[w_j = x_j, j \in Pre^i(\mathcal{O})]}) \right]. \tag{2.7}$$

$\pi(N)$: Representa o conjunto de todas as permutações possíveis das n variáveis explicativas.

$\mathcal{O}$: Uma permutação específica dentro de $\pi(N)$, correspondendo a uma ordem particular de entrada das variáveis explicativas. Cada $\mathcal{O}$ define uma sequência única para avaliar a contribuição marginal de cada variável.

$Pre^i(\mathcal{O})$: Conjunto de índices que precedem a variável explicativa i na permutação $\mathcal{O}$.

w: Instância de referência amostrada do espaço de variáveis explicativas $\mathcal{X}$, servindo como base para a perturbação controlada das variáveis não incluídas no conjunto sob análise.

$p(w)$: Função de probabilidade que governa a amostragem de w. O método assume independência entre variáveis explicativas, simplificando $p(w)$ para o produto das distribuições marginais.

$w_{[w_j = x_j, j \in Q]}$: Notação para composição de instâncias de intervenção, em que as variáveis explicativas no conjunto Q mantêm os valores da instância original x, enquanto as demais variáveis assumem os valores da instância de referência w.

$f(\cdot)$: Modelo de predição sendo explicado, tratado como função caixa-preta (*black box*), cujas saídas são observadas frente às perturbações nas entradas.

A implementação prática do método, apresentada no Algoritmo 1 conforme Strumbelj e Kononenko (2014), utiliza o seguinte estimador, derivado da formulação teórica da Equação (2.7):

$$\hat{\varphi} = \frac{1}{m} \sum_{j=1}^{m} V_j, \tag{2.8}$$

em que,

$$V_j = f(\vec{b}_1) - f(\vec{b}_2). \tag{2.9}$$

O estimador $\hat{\varphi}_i$ representa a aproximação amostral do verdadeiro valor de Shapley $\varphi_i(x)$. O parâmetro m determina o número de amostras utilizadas na aproximação, funcionando como um controlador do compromisso entre a precisão da estimativa e o custo computacional envolvido. Cada V_j corresponde a uma variável aleatória que representa uma amostra individual da diferença marginal, sendo cada V_j uma realização concreta do termo entre colchetes na Fórmula (2.7).

As instâncias $\vec{b}_1$ e $\vec{b}_2$ são versões modificadas dos dados criadas para medir o efeito de uma variável i. Em $\vec{b}_1$, a variável i e todas as que a antecedem na ordem de análise assumem os valores da instância original x, enquanto as demais variáveis provêm de uma instância de referência w. Já $\vec{b}_2$ é idêntico, exceto pela exclusão da variável i, mantendo apenas as anteriores a ela com os valores de x. A diferença entre os resultados do modelo para $\vec{b}_1$ e $\vec{b}_2$ isola exatamente a contribuição marginal da i-ésima variável.

3 Resultados

Este trabalho apresenta, como resultado inicial, o agrupamento via K-means funcional das curvas mensais dos ciclos de produção agrícola, de acordo com as fases do fenômeno ENOS (El Niño, La Niña e Neutro).

A Figura 1 mostra o resultado da convergência do processo de atribuição das 58 funções do ciclo de produção agrícola aos respectivos grupos no Estado de Goiás. Cada função é definida ao longo de um intervalo de oito meses, compreendendo o período de outubro a maio, conforme descrito na Seção 2.1.

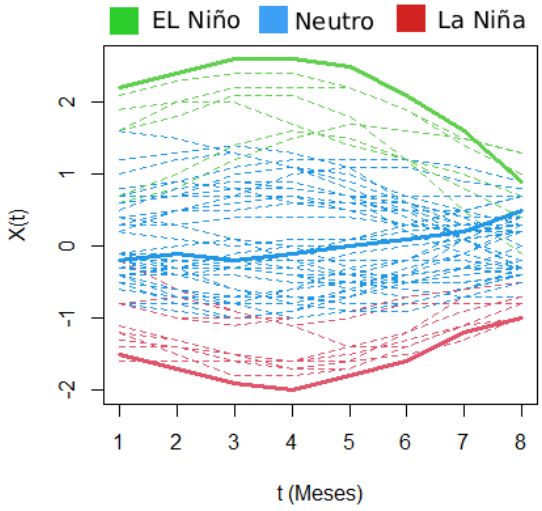

Figura 1 – Agrupamento funcional dos ciclos de produção agrícola em clusters associados às três fases do ENOS, no Estado de Goiás (1961–2019).

Fonte: Elaborado pelo autor.

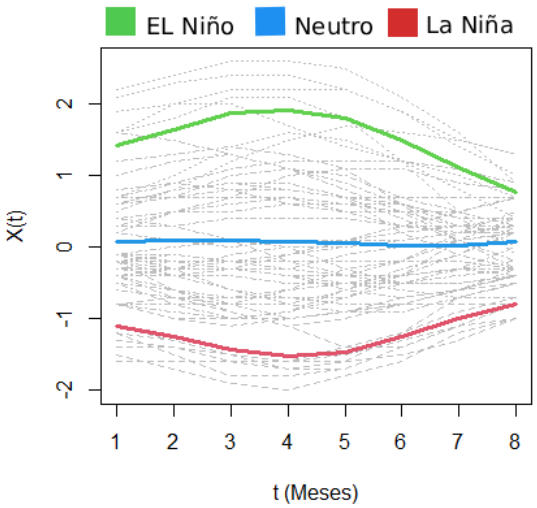

Figura 2 – Centróides funcionais resultantes da análise de K-means funcional no Estado de Goiás (1961–2019).

Fonte: Elaborado pelo autor.

As curvas individuais de cada cluster, destacadas por coloração mais intensa na Figura 1, têm como objetivo facilitar a visualização das regiões correspondentes a cada agrupamento. Observa-se, no resultado da atribuição, um total de 58 funções que representam os ciclos de produção agrícola: 7 para o cluster El Niño, 11 para o La Niña e 40 para o Neutro, considerando a avaliação geral referente aos anos de 1961 a 2019. No período de 1990 a 2019, foco deste estudo, foram identificados 29 ciclos, dos quais 4 foram atribuídos ao cluster El Niño, 6 ao La Niña e 19 ao Neutro. A Figura 2, ao lado, apresenta os centróides finais, que correspondem à média funcional das curvas pertencentes a cada cluster.

A Figura 3 mostra o resultado da convergência do processo de atribuição das 58 funções

de ciclo de produção agrícola aos respectivos grupos para o Estado do Rio Grande do Sul. Cada função é definida ao longo de um intervalo de dez meses, compreendendo o período de setembro a junho. A Figura 4, por sua vez, apresenta os centróides funcionais de cada cluster, os quais também correspondem à média das funções contidas em seus respectivos agrupamentos.

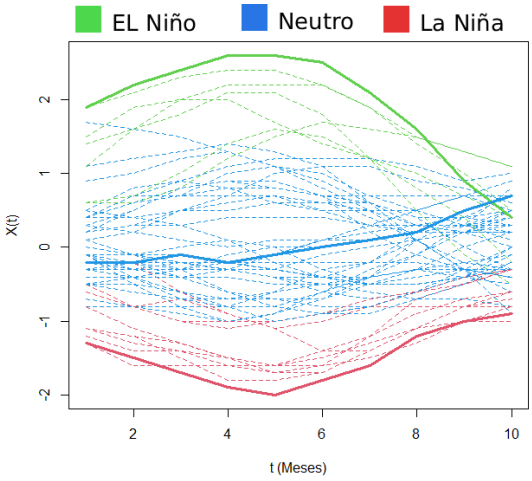

Figura 3 – Agrupamento funcional dos ciclos de produção agrícola em clusters associados às três fases do ENOS, no Estado do Rio Grande do Sul (1961–2019)

Fonte: Elaborado pelo autor.

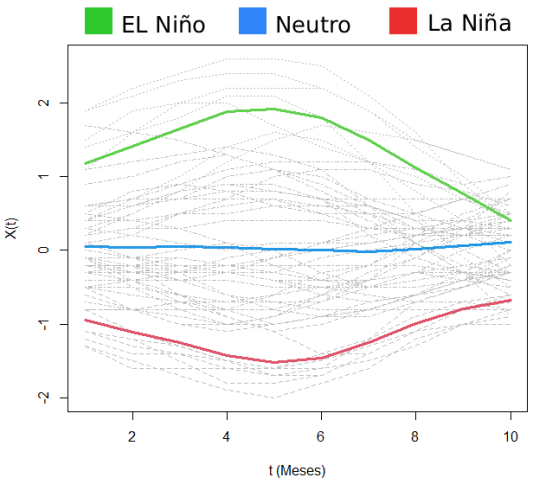

Figura 4 – Centróides funcionais resultantes da análise de K-means funcional no Estado do Rio Grande do Sul (1961–2019)

Fonte: Elaborado pelo autor.

Na Figura 3, no Estado do Rio Grande do Sul, das 58 funções dos ciclos de produção agrícola, 7 estão associadas ao cluster El Niño, 11 ao La Niña e 40 ao Neutro. Foram considerados, neste trabalho, 29 ciclos de produção referentes ao período de interesse (1990–2019), dos quais 6 pertencem ao cluster La Niña, 3 ao El Niño e 20 ao Neutro.

Com os clusters determinados, foi realizada a análise de variância funcional (FANOVA), com o nível de significância estabelecido em 5%, e as hipóteses de interesse testadas estão na seção (1.1).

A Tabela 1, a seguir, apresenta os resultados obtidos pelo teste de hipótese FP da análise de variância funcional (FANOVA), em testes individuais por variável climática, em escala semanal por meio da comparação dos clusters obtidos pelo K-means. Foram avaliadas as funções de observações semanais entre setembro e junho, que delimitam o ciclo de produção agrícola neste estudo.

Tabela 1 – Valores de *p* do teste FP da FANOVA comparando os efeitos das fases do ENOS sobre as variáveis climáticas por Estado (1990–2019)

Variáveis	Rio Grande do Sul	Goiás
Temperatura Máxima	0,008	0,273
Temperatura Mínima	0,000	0,642
Precipitação Acumulada	0,000	1,000

Fonte: Adaptado de INMET (2019).

Com o nível de significância adotado de $\alpha = 5\%$, conforme apresentado na Tabela 1, foram analisadas as variáveis climáticas: temperatura máxima, temperatura mínima e precipitação acumulada. Para o Estado de Goiás, observa-se que os valores de *p* encontram-se, em geral, acima de 0,05, o que permite concluir que não há evidências suficientes para rejeitar a hipótese nula. Portanto, nesse Estado, os efeitos das três fases do fenômeno ENOS sobre as variáveis climáticas podem ser considerados iguais. Por outro lado, para o Estado do Rio Grande do Sul, os valores de *p* estão abaixo do nível α, indicando evidências para rejeitar a hipótese nula (1.1). Nesse caso, os resultados sugerem a existência de diferenças significativas entre as fases do fenômeno ENOS com base nos dados analisados.

Na Figura 5, a seguir, são exibidas as curvas médias funcionais, ao longo de 43 semanas, por variável climática (A, B e C) para o Estado do Rio Grande do Sul, condicionadas ao resultado da Tabela 1. Conforme mencionado anteriormente, há diferenças significativas, em média, entre essas curvas por cluster do fenômeno ENOS.

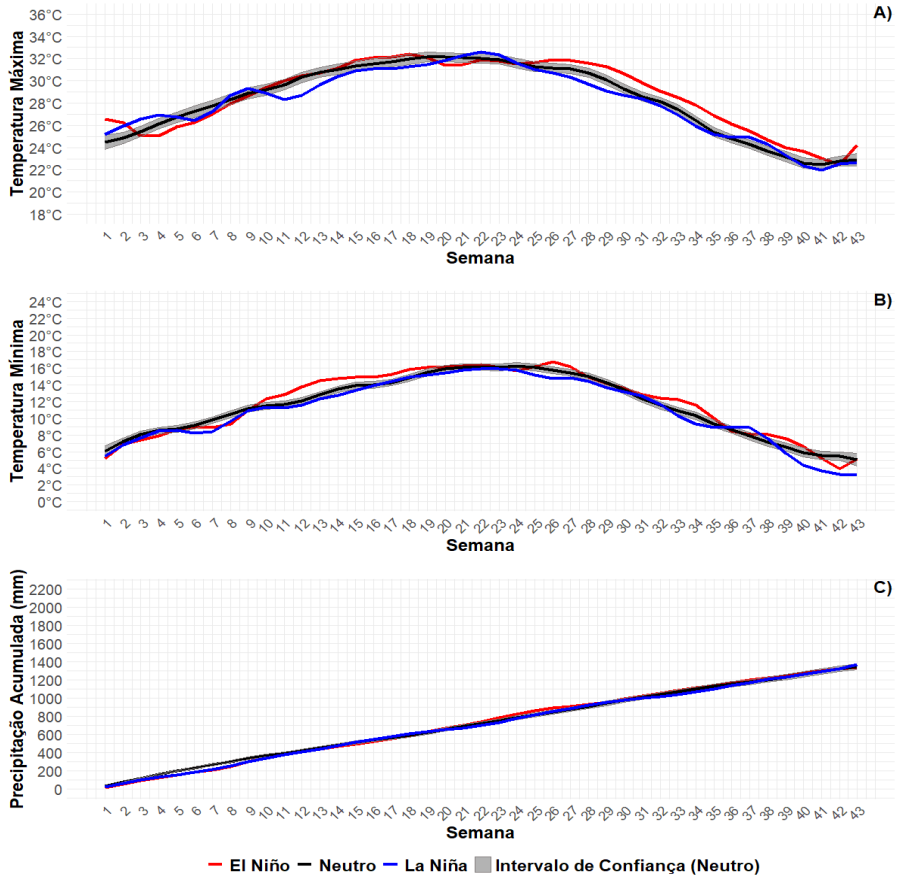

Figura 5 – Curva média funcional geral e intervalos de confiança de 95% (período 1990–2019) em torno da média funcional do neutro para o Estado do Rio Grande do Sul, por fase do ENOS, para as variáveis climáticas: A) temperatura máxima, B) temperatura mínima e C) precipitação acumulada.

Fonte: Elaborado pelo autor.

Com a descrição das curvas médias funcionais ao longo das semanas do ciclo de produção agrícola, é notável uma inclinação geral na tendência por fases do fenômeno, conforme apresentado na Figura 5 A) temperatura máxima e na Figura 5 B) temperatura mínima, com algumas semanas em torno da semana 22, que se estende de 26 de janeiro a 1º de fevereiro, apresentando as maiores temperaturas médias neste ciclo, no Rio Grande do Sul. A Figura 5 C) mostra que o ciclo de produção agrícola acumula, em média, por fase do ENOS, entre 1200 mm e 1400 mm de precipitação. No entanto, há variações entre as curvas. Assim, torna-se de interesse identificar quais fases do ENOS se diferem, em média, para cada variável climática. A abordagem adotada consistiu na realização de testes FP da FANOVA, dois a dois, para todas as combinações distintas possíveis, conforme descrito na Seção 1.4.

Além disso, para o Estado de Goiás, onde não houve evidências significativas para rejeitar H_0, ou seja, não há motivos para considerar médias funcionais distintas por fase do fenômeno ENOS, verifica-se que, ao nível de 5% de significância, as curvas médias ao longo de 34 semanas são iguais. Não foram identificados, nos dados, efeitos significativos do ENOS sobre as médias

das variáveis climáticas (Figura 6, A, B e C), considerando-se, portanto, uma única média funcional geral, como mostrado na Figura 6, abaixo.

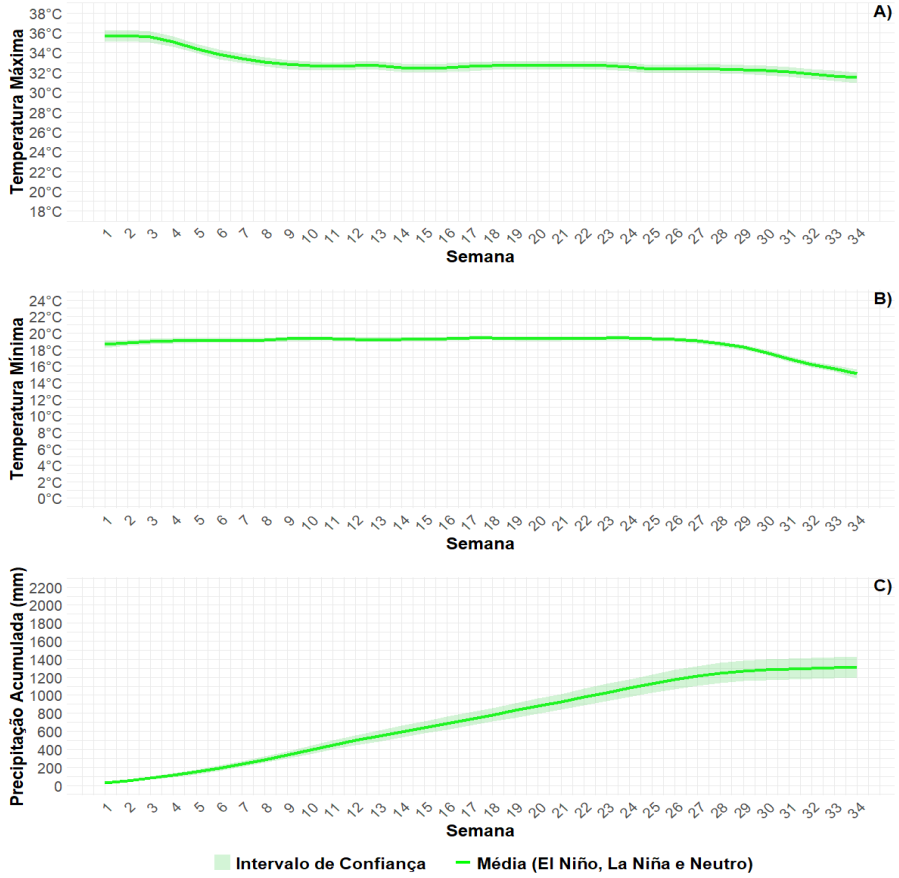

Figura 6 – Curva média funcional geral e intervalos de confiança de 95% (período 1990–2019) em torno da média funcional para o Estado de Goiás, por fase do ENOS, para as variáveis climáticas: A) temperatura máxima, B) temperatura mínima e C) precipitação acumulada.

Fonte: Elaborado pelo autor.

Como mostrado na Figura 6, o Estado de Goiás apresenta um padrão mais estável das temperaturas semanais, diferindo do observado no Estado do Rio Grande do Sul. Conforme os resultados anteriores, para o Estado de Goiás não houve evidências para a rejeição da hipótese nula, o que implica na descontinuidade da avaliação da importância dessas variáveis, já que não foram identificados indícios do efeito médio do ENOS.

Prosseguindo com o Estado do Rio Grande do Sul, as comparações múltiplas envolvem o problema de inflação do nível de significância α, que é a taxa de erro tipo I que deve ser controlada. Dessa forma, a Tabela 2 apresenta os valores p ajustados pelo método de Holm–Bonferroni (2.4), visando esse controle

Tabela 2 – Valores *p* dos testes múltiplos FP (FANOVA) com correção Holm-Bonferroni para o Rio Grande do Sul (1990-2019)

Variáveis	**Hipóteses**		
	H_0: El Niño = Neutro vs H_1: El Niño $\neq$ Neutro	H_0: El Niño = La Niña vs H_1: El Niño $\neq$ La Niña	H_0: La Niña = Neutro vs H_1: La Niña $\neq$ Neutro
Temp. Máxima	0,002	0,260	0,006
Temp. Mínima	0,002	0,000	0,000
Precip. Acumulada	0,000	0,000	0,000

Fonte: Elaborado pelo autor.

Os valores *p* obtidos na Tabela 2, ao nível de 5% de significância, indicam que, para a temperatura máxima, o valor *p* é aproximadamente 0,260, acima de 0.05, para a hipótese de comparação El Niño = La Niña. Nesse caso, não há evidências para rejeitar essa hipótese, o que implica assumir que as curvas de temperatura máxima nas fases El Niño e La Niña são iguais, em média. Para as demais hipóteses de comparação par a par entre as variáveis climáticas, estas foram rejeitadas ao mesmo nível de significância. Dessa forma, há evidências para assumirmos que as curvas médias das variáveis climáticas por grupo do fenômeno ENOS são, em média, distintas.

Através desses resultados, variáveis categóricas foram criadas como mostra o Quadro 1, para refletirem as evidências encontradas no FANOVA dois a dois no Estado do Rio Grande do Sul, onde esses efeitos do fenômeno ENOS são significativos.

Quadro 1 – Variáveis categóricas com base nos testes de hipóteses (1990-2019)

Variável	**Categoria**	**Fenômenos Agrupados**
Temperatura Máxima	Tmax_N	Neutro
	Tmax_EL_LA	El Niño e La Niña
Temperatura Mínima	Tmin_N	Neutro
	Tmin_EL	El Niño
	Tmin_LA	La Niña
Precipitação Acumulada	P_N	Neutro
	P_EL	El Niño
	P_LA	La Niña

Fonte: Elaborado pelo autor.

A coluna “Categoria” no Quadro 1 apresenta como essas variáveis formam categorias definidas por ano, representando a etapa final do ciclo produtivo agrícola, que é a colheita, realizada no ano seguinte, logo após o plantio e o desenvolvimento da cultura. Portanto, os dados de rendimento devem compreender o período entre 1991 e 2019.

A Figura 7 representa um caso do uso do modelo de regressão local, cujo objetivo é capturar a curva da tendência tecnológica dos dados de rendimento de arroz no município de Alegrete. Sua flexibilidade em realizar ajustes locais para estimar a curva foi um grande avanço, pois, em muitos casos, o comportamento dos dados de rendimento pode ser não linear.

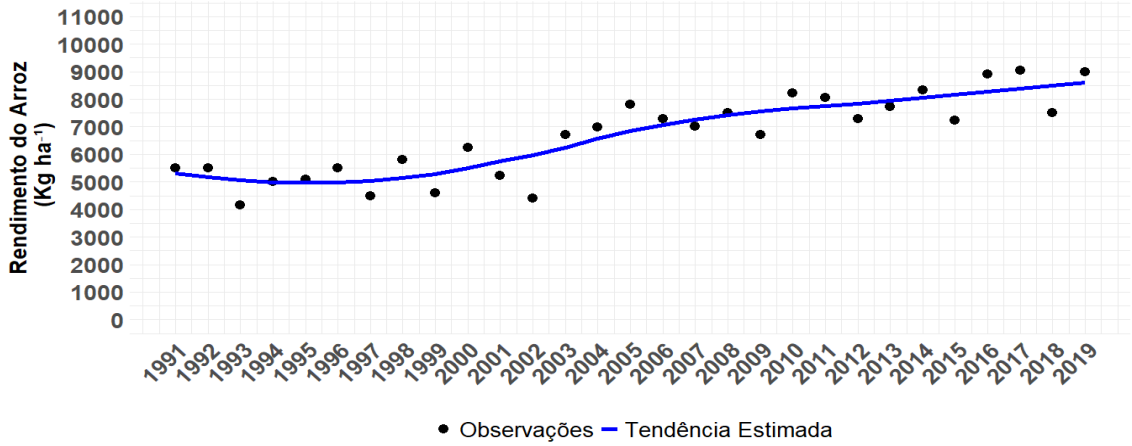

Figura 7 – Ajuste da tendência evolutiva tecnológica aos dados de rendimento do arroz no município de Alegrete-RS (1991–2019)

Fonte: Elaborado pelo autor.

No total, foram ajustados 415 modelos de regressão local para os municípios do Rio Grande do Sul, distribuídos em três culturas. Para o arroz, foram 68 municípios, para o feijão, 182 e para a soja, 165. Na Figura 8, já é possível observar que os dados estão estáveis, resultado das etapas de cálculo do desvio relativo apresentadas em 2.5 e do ajuste do rendimento, tomando o ano de 2019 como base, conforme 2.6.

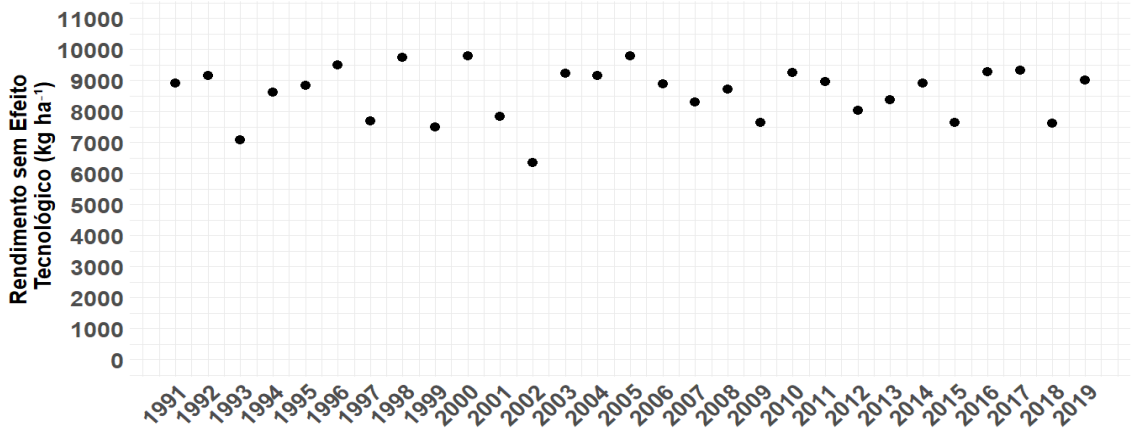

Figura 8 – Rendimento da produção de arroz no município de Alegrete-RS (1991-2019)

Fonte: Elaborado pelo autor.

Dessa forma, a Figura 8 apresenta o resultado final do ajuste obtido com o modelo de regressão local. O método mostrou-se capaz de capturar a tendência dos dados e de adaptar-se

às flutuações locais. Assim, os dados de rendimento foram corrigidos quanto à influência da evolução tecnológica ao longo dos anos e projetados para um cenário atual, preservando as variações climáticas que constituem o foco deste estudo.

Em seguida, com os dados de rendimento de todos os municípios nivelados com base no cenário climático do ano de 2019, a etapa seguinte consistiu em utilizar os dados de rendimento de cada município, agrupados por cultura ao longo dos anos (1991–2019). A abordagem de agrupamento em duas classes de rendimento (alta e baixa) visa identificar padrões e comportamentos semelhantes entre os dados, de acordo com sua magnitude. A Tabela 3 apresenta as estatísticas descritivas dos grupos de rendimento alto e baixo por cultura, com base no resultado do agrupamento obtido pelo método k-means.

Tabela 3 – Estatísticas descritivas de rendimento kg ha^{-1} para municípios do Rio Grande do Sul (1991-2019)

	Grupo de Alto Rendimento				**Grupo de Baixo Rendimento**			
Culturas	**Mín.**	**Média**	**Máx.**	**D.P.**	**Mín.**	**Média**	**Máx.**	**D.P.**
Arroz	7175	8130	11020	704	727	6216	7170	1179
Feijão	1469	1928	5731	485	108	1007	1467	287
Soja	2739	3511	7879	582	222	1965	2737	569

Fonte: Elaborado pelo autor.
Nota: D.P. = Desvio Padrão.

Conforme, observado na Tabela 3, o arroz apresenta a maior média de rendimento (em kg ha^{-1}) entre os municípios pertencentes ao grupo de alta produtividade. A soja ocupa a segunda posição em termos de escala de produção (kg ha^{-1}), enquanto o feijão é produzido em quantidades menores nos municípios avaliados do Estado do Rio Grande do Sul. Ao observar, por cultura, os valores mínimos do grupo de alto rendimento e os máximos do grupo de baixo rendimento, nota-se que não houve sobreposição de valores, o que indica que a divisão realizada pelo algoritmo do método k-means convergiu corretamente as observações aos grupos.

Dando continuidade, no resultado anterior as observações dos rendimentos das culturas foram categorizadas de forma binária com base no grupo a que pertencem. Se a observação estiver no grupo de alto rendimento, ela recebe valor 1, caso contrário, se estiver no grupo de baixo rendimento, recebe valor 0. Dessa forma, obtém-se um conjunto de dados completamente categórico para ser utilizado no modelo de classificação supervisionado "Random Forest". Esse modelo foi empregado com o objetivo de aproveitar sua estrutura robusta para avaliar a contribuição das variáveis na predição média do modelo com base no conjunto de treino. Essa contribuição é mensurada por meio dos valores de Shapley.

Os dados são compostos por uma variável resposta binária, indicando rendimento alto ou baixo, e por variáveis preditoras, entre as quais alguns municípios do Estado do Rio Grande do Sul são utilizados para incorporar informações adicionais e melhorar a acurácia do modelo. Para

a cultura do arroz são 68 municípios, para o feijão são 182 e para a soja 165, além de temperatura máxima, temperatura mínima e precipitação acumulada, conforme apresentado no Quadro 1.

Na Tabela 4 são apresentados os resultados obtidos a partir da execução de três modelos de classificação Random Forest, um para cada cultura, desenvolvidos segundo os procedimentos de validação cruzada e ajuste de hiperparâmetros, que são práticas fundamentais em aprendizagem de máquina para garantir a robustez e a capacidade de generalização dos modelos. A validação cruzada permite avaliar o desempenho do modelo de forma mais confiável, reduzindo o viés associado a uma única divisão dos dados. Foram consideradas cinco partições, enquanto o ajuste de hiperparâmetros buscou identificar a configuração ideal do algoritmo com base nos intervalos de busca definidos na grade.

Tabela 4 – Desempenho do treinamento do modelo Random Forest na classificação binária de rendimento alto e baixo (1991–2019)

Cultura	**Acurácia**
Arroz	79,8%
Feijão	78,1%
Soja	76,5%

Fonte: Elaborado pelo autor.

Como resultado, na Tabela 4, observam-se acurácias das três culturas próximas de 80%. Esse é um bom desempenho, considerando que o número de variáveis explicativas é igual a quatro e que há complexidade associada à presença de multicolinearidade. Assim, o modelo mostrou-se uma alternativa adequada diante do desempenho obtido no treinamento. O intuito de utilizar esse modelo está em explorar sua estrutura para lidar com a natureza dos dados e identificar padrões de importância com base no processo de treino, e não em utilizá-lo para fins de previsão.

A aplicação dos valores SHAP ao modelo de Random Forest permite quantificar a importância das variáveis por meio da média dos valores absolutos de suas contribuições locais para todas as instâncias do conjunto de dados nas quais foram normalizadas para porcentagem. Essa abordagem preserva a propriedade fundamental de eficiência dos valores de Shapley, garantindo que, para qualquer previsão individual, a soma das contribuições atribuídas a cada variável seja exatamente igual à diferença entre a saída do modelo para aquela instância (Strumbelj e Kononenko, 2014). Como a variável resposta é binária, o modelo gera duas probabilidades, sendo uma correspondente à classe de alto rendimento, adotada como referência nesta análise, e outra referente à classe de baixo rendimento, que é o seu complemento da probabilidade da classe de alto rendimento.

Na Figura 9 apresenta-se o resultado final da análise de importância das variáveis climáticas em escala semanal, associadas ao efeito do ENOS e aos municípios. Essa importância é

representada pela proporção da contribuição média global de cada variável explicativa para a classificação da classe de alto rendimento do arroz no modelo Random Forest.

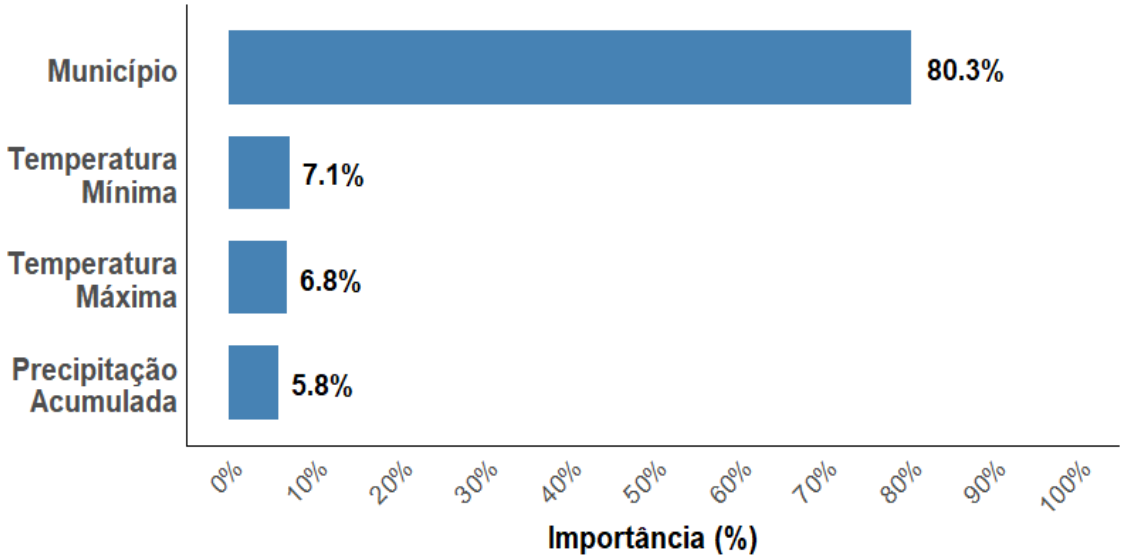

Figura 9 – Proporção de importância das variáveis para as classes de rendimento do arroz no Estado do Rio Grande do Sul para os anos 1991-2019
Fonte: Elaborado pelo autor.

Como mostrado na Figura 9, o município apresenta importância expressiva, com 80,3%, representando a maior parcela da magnitude média das contribuições locais para as classes de alto rendimento do arroz. Esse resultado indica que a variável regional exerce papel determinante na produtividade do cultivo. Em seguida, observa-se que a temperatura mínima contribui com aproximadamente 7,1%, valor próximo ao da temperatura máxima, com 6,8%, representando uma fração pequena, porém relevante. Por fim, a precipitação acumulada apresenta 5,8% de importância média absoluta na previsão do modelo. Essas variáveis climáticas foram construídas com base nas evidências do fenômeno ENOS, obtidas pelo teste de hipóteses FANOVA aplicado aos ciclos entre 1990 e 2019 no Estado do Rio Grande do Sul.

Em relação à Figura 10, são apresentadas as proporções de importância das mesmas variáveis resultantes da FANOVA para o feijão, abrangendo um número de municípios maior do que na cultura do arroz.

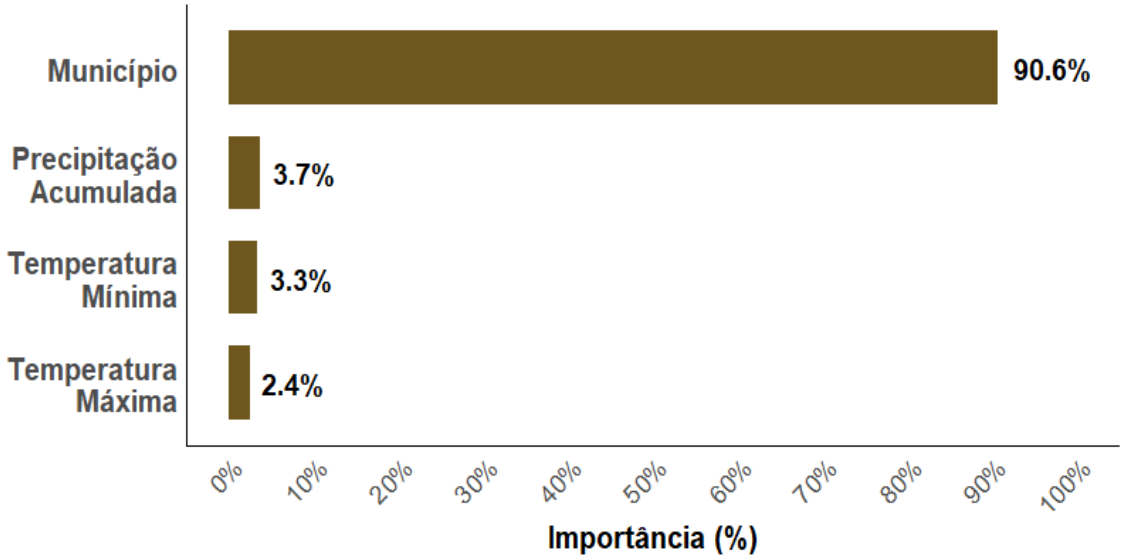

Figura 10 – Proporção de importância das variáveis para as classes de rendimento do feijão no Estado do Rio Grande do Sul para os anos 1991-2019
Fonte: Elaborado pelo autor.

É possível observar na Figura 10 que o município apresenta a maior fração individual, com 90,6% de importância, com base na média absoluta das contribuições das instâncias municipais para a previsão média do modelo na classe de alto rendimento do feijão. Esse resultado evidencia a relevância das produções regionais como fator decisivo para a cultura. Em seguida, a precipitação acumulada apresenta 3,7% e a temperatura mínima 3,3% de importância, indicando, com base no resultado da FANOVA, que o comportamento médio distinto entre as fases do fenômeno ENOS, reflete uma magnitude de importância baixa em comparação ao município. A temperatura máxima, por sua vez, apresenta 2,4% de importância, sendo a menor entre as variáveis analisadas, em que os efeitos médios das fases foram semelhantes apenas para os eventos El Niño e La Niña.

Por fim, na Figura 11, referente à cultura da soja, são apresentadas também as porcentagens de importância individual das variáveis meteorológicas e do município para o modelo de classificação de rendimento Random Forest.

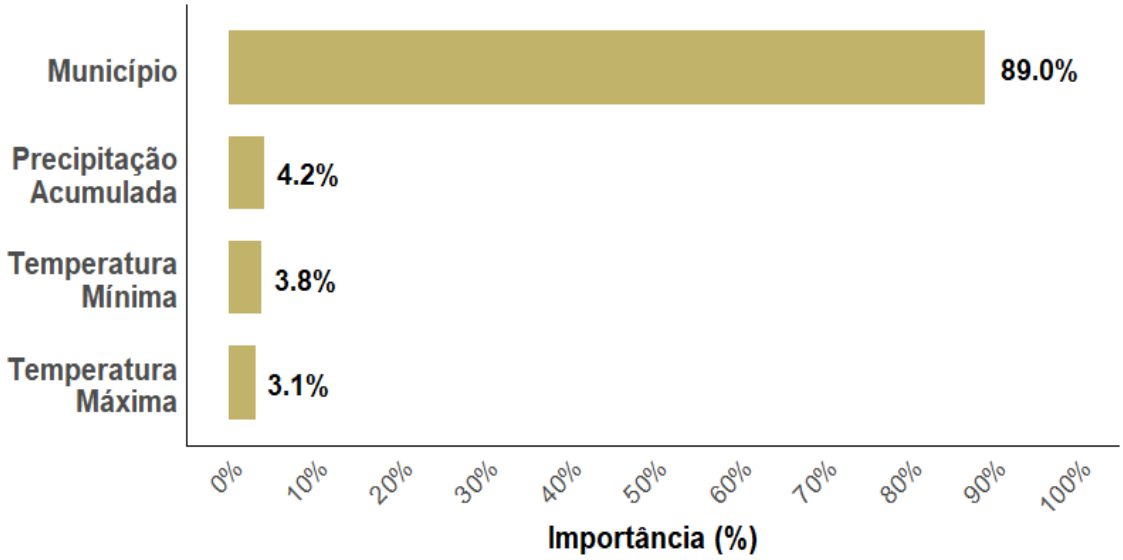

Figura 11 – Proporção de importância das variáveis para as classes de rendimento da soja no Estado do Rio Grande do Sul para os anos 1991-2019
Fonte: Elaborado pelo autor.

Os resultados obtidos pelos valores SHAP para o modelo Random Forest aplicado à soja indicaram uma magnitude de aproximadamente 89,0% de importância relativa individual dos municípios em relação ao alto rendimento da cultura. Esse resultado pode sinalizar, a princípio, fatores mais determinísticos associados às características municipais, como tipo de solo, altitude e outros aspectos que podem ser investigados de forma mais detalhada, tornando essa variável de alto potencial para a previsão média da classificação da soja, conforme também observado nas culturas anteriores.

A precipitação acumulada apresentou importância reduzida, de 4,2%, seguida da temperatura mínima, com 3,8% de importância, também pequena, na qual houve evidência de diferenças significativas entre as fases do ENOS. Já a temperatura máxima apresentou a menor proporção de importância, de 3,1%, em que os efeitos médios das fases La Niña e El Niño mantiveram-se semelhantes, segundo as evidências obtidas pela FANOVA. Isso indica que, no modelo, os efeitos do ENOS sobre essa variável não exerceram contribuição expressiva para a previsão média absoluta das classes de alto rendimento.

Conclusão

O presente trabalho teve como objetivo investigar a influência do fenômeno El Niño Oscilação Sul (ENOS) sobre os ciclos de produção agrícola, com base em variáveis climáticas em escala semanal, como temperatura máxima, temperatura mínima e precipitação acumulada, considerando especificamente os estados de Goiás e Rio Grande do Sul. Foram encontradas evidências significativas no teste de hipótese FANOVA, com o Estado de Goiás apresentando estabilidade, sem diferenças médias relevantes nas variáveis climáticas entre as fases do fenômeno ENOS. Por outro lado, no Estado do Rio Grande do Sul observaram-se evidências de diferenças entre as fases do ENOS. Assim, buscou-se analisar como os efeitos médios dessas fases, identificados nas variáveis climáticas, contribuem, juntamente com o fator municipal, para a classificação das produtividades de arroz, feijão e soja. A abordagem metodológica empregada integrou técnicas de análise estatística clássica, análise funcional e aprendizado de máquina. Em síntese, os métodos aplicados, K-means funcional, Análise de Variância Funcional (FANOVA), Regressão Local, K-means clássico e modelo Random Forest, permitiram uma compreensão abrangente dos efeitos do ENOS sobre as variáveis climáticas e produtivas.

A aplicação do modelo Random Forest aos dados que representam apenas variações climáticas ajustadas para o ano de 2019 é complementada pela análise dos valores SHAP, o que possibilitou identificar a relevância individual das variáveis climáticas e dos municípios sobre o alto rendimento agrícola. Para o arroz, o feijão e a soja, a variável categórica município destacou-se como o fator mais determinante em termos de importância preditiva média absoluta na classificação das classes de alto rendimento das três culturas. Essa importância observada para os dados avaliados entre 1991 e 2019 indica que as variações regionais expressas pela variável município possivelmente estão associadas a características de solo e relevo, que podem ser investigadas em trabalhos futuros.

Vale ressaltar que as variáveis climáticas expressas em categorias refletem as influências das fases do ENOS, consideradas significativas segundo a FANOVA no Estado do Rio Grande do Sul, e representam uma pequena porcentagem de importância para o alto rendimento das culturas. Em ordem decrescente, as variáveis climáticas das culturas de feijão e soja, ambas pertencentes à família Leguminosae, apresentam a mesma hierarquia de importância mostrada nas Figuras 10 e 11. A temperatura máxima foi a variável que menos contribuiu para a previsão da importância média absoluta do modelo, sendo que os efeitos de La Niña e El Niño mostraram-se equivalentes a 5% de significância nos ciclos de produção agrícola entre 1990 e 2019 para essas duas culturas, conforme o modelo Random Forest. Para o arroz, a variável que apresentou a menor porcentagem de importância nos valores SHAP foi a precipitação acumulada. Contudo, embora o fenômeno ENSO apresente significância estatística na média funcional das variáveis climáticas do Estado do Rio Grande do Sul, sua contribuição, descrita pelos valores SHAP, é baixa no modelo de

classificação Random Forest. Esse resultado indica que, neste estudo, o alto rendimento das culturas é mais influenciado pela localização geográfica.

Dessa forma, os resultados obtidos reforçam a importância de avaliar os efeitos do ENOS sobre o clima e a produção agrícola. A fim de compreender e subsidiar a adoção de estratégias adaptativas e de políticas públicas voltadas a sistemas agrícolas sensíveis às oscilações climáticas induzidas pelo fenômeno, contribuindo, assim, para uma agricultura mais sustentável.

Referências

ABDI, H. Holm's sequential bonferroni procedure. **Encyclopedia of research design**, Thousand Oaks, California, v. 1, n. 8, p. 1–8, 2010. Citado na página 19.

AGRAWAL, T. **Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient**. 1st. ed. New York, NY: Apress, 2021. Disponível em: <https://doi.org/10.1007/978-1-4842-6579-6>. Citado 2 vezes nas páginas 21 e 30.

ANDERSON, W. *et al.* Trans-pacific enso teleconnections pose a correlated risk to agriculture. **Agricultural and Forest Meteorology**, v. 262, p. 298–309, 2018. ISSN 0168-1923. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168192318302454>. Citado na página 14.

BJERKNES, J. Atmospheric teleconnections from the equatorial pacific. **Monthly weather review**, v. 97, n. 3, p. 163–172, 1969. Citado na página 16.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 30.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 29.

CAI, W. *et al.* Climate impacts of the el niño–southern oscillation on south america. **Nature Reviews Earth & Environment**, v. 1, n. 4, p. 215–231, abr. 2020. ISSN 2662-138X. Disponível em: <https://doi.org/10.1038/s43017-020-0040-3>. Citado na página 14.

CHOWDHURY, S. *et al.* Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance. In: IEEE. **2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)**. [S.l.], 2022. p. 17–25. Citado na página 21.

CLEVELAND, W. S. Robust locally weighted regression and smoothing scatterplots. **Journal of the American statistical association**, Taylor & Francis, v. 74, n. 368, p. 829–836, 1979. Citado na página 20.

CONAB. **Calendário de plantio e colheita de grãos no Brasil 2022**. Brasília: Conab, 2022. Disponível em: <https://www.gov.br/conab/pt-br/acesso-a-informacao/institucional/publicacoes/arquivos-de-paginas/calendariozplantiozezcolheitazjunz2022.pdf>. Citado 2 vezes nas páginas 14 e 23.

CONAB. **Acompanhamento da safra brasileira de grãos, 10º levantamento – safra 2024/25**. 2025. Acesso em: 07 ago. 2025. Disponível em: <https://www.gov.br/conab/pt-br/atuacao/informacoes-agropecuarias/safras/safra-de-graos>. Citado na página 14.

COSTA-NETO, G. *et al.* Environmental clusters defining breeding zones for tropical irrigated rice in brazil. **Agronomy Journal**, v. 116, n. 3, 2024. Disponível em: <https://acsess.onlinelibrary.wiley.com/doi/abs/10.1002/agj2.21481>. Citado na página 18.

EMBRAPA. **Produção dos principais grãos**. 2021. Acesso em: 14 out. 2025. Disponível em: <https://www.embrapa.br/visao-de-futuro/trajetoria-do-agro/desempenho-recente-do-agro/principais-graos>. Citado na página 20.

FEBRERO-BANDE, M.; FUENTE, M. O. D. L. Statistical computing in functional data analysis: The r package fda. usc. **Journal of statistical Software**, v. 51, p. 1–28, 2012. Citado na página 18.

FERRATY, F.; VIEU, P. **Nonparametric functional data analysis: theory and practice**. [S.l.]: Springer, 2006. 23-24 p. Citado na página 17.

GIJBELS, I.; PROSDOCIMI, I. Loess. **WIREs Computational Statistics**, v. 2, n. 5, p. 590–599, 2010. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.104>. Citado 2 vezes nas páginas 20 e 27.

GÓRECKI, T.; SMAGA Łukasz. A comparison of tests for the one-way anova problem for functional data. **Computational Statistics**, v. 30, n. 4, p. 987–1010, 2015. ISSN 1613-9658. Disponível em: <https://doi.org/10.1007/s00180-015-0555-0>. Citado 2 vezes nas páginas 19 e 25.

HARTIGAN, J. A.; WONG, M. A. A k-means clustering algorithm. **Journal of the Royal Statistical Society Series C**, v. 28, n. 1, p. 100–108, 1979. Disponível em: <https://EconPapers.repec.org/RePEc:bla:jorssc:v:28:y:1979:i:1:p:100-108>. Citado 2 vezes nas páginas 17 e 18.

HASTIE, T. *et al.* **The elements of statistical learning**. [S.l.]: Springer series in statistics New-York, 2009. Citado na página 21.

HEINEMANN, A. B.; SENTELHAS, P. C. Environmental group identification for upland rice production in central brazil. **Scientia Agricola**, Escola Superior de Agricultura "Luiz de Queiroz", v. 68, n. 5, p. 540–547, Sep 2011. ISSN 0103-9016. Disponível em: <https://doi.org/10.1590/S0103-90162011000500005>. Citado na página 28.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian journal of statistics**, JSTOR, p. 65–70, 1979. Citado 3 vezes nas páginas 20, 26 e 27.

IBGE. **Produção Agropecuária no Brasil**. 2024. Disponível em: <https://www.ibge.gov.br/explica/producao-agropecuaria/>. Acesso em: 21 ago. 2025. Citado na página 14.

IBGE. **Tabela 1612 - Produto Interno Bruto a preços correntes**. 2024. Sistema IBGE de Recuperação Automática - SIDRA. Disponível em: <https://sidra.ibge.gov.br/tabela/1612>. Citado na página 23.

IIZUMI, T. *et al.* Impacts of el niño southern oscillation on the global yields of major crops. **Nature Communications**, v. 5, n. 1, p. 3712, 2014. ISSN 2041-1723. Disponível em: <https://doi.org/10.1038/ncomms4712>. Citado na página 16.

INMET. **Instituto Nacional de Meteorologia: Dados Históricos de Clima**. 2019. Acessado em: 02 ago. 2025. Disponível em: <https://portal.inmet.gov.br/dadoshistoricos>. Citado na página 23.

JAMES, G. *et al.* **An introduction to statistical learning: with applications in R**. [S.l.]: Springer, 2013. v. 103. Citado 2 vezes nas páginas 21 e 30.

KREYSZIG, E. **Introductory Functional Analysis with Applications**. Wiley, 1991. (Wiley Classics Library). ISBN 9780471504597. Disponível em: <https://books.google.com.br/books?id=AQtMEAAAQBAJ>. Citado na página 17.

LLOYD, S. Least squares quantization in pcm. **IEEE transactions on information theory**, IEEE, v. 28, n. 2, p. 129–137, 1982. Citado na página 17.

MARENGO, J. A. *et al.* O maior desastre climático do brasil: chuvas e inundações no estado do rio grande do sul em abril-maio 2024. **Estudos Avançados**, Instituto de Estudos Avançados da Universidade de São Paulo, v. 38, n. 112, p. 203–228, 2024. ISSN 0103-4014. Disponível em: <https://doi.org/10.1590/s0103-4014.202438112.012>. Citado na página 14.

NOAA. **Historical ENSO episodes (1950–present): cold and warm episodes by season**. 2020. <https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php>. National Weather Service, Climate Prediction Center. Acesso em: 10 out. 2025. Citado na página 23.

OROZCO, N.; ORTIZ, S.; OSPINA-TASCÓN, G. A. Functional data analysis applications in medicine: A systematic review. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, v. 17, n. 2, p. e70026, 2025. Citado na página 18.

PENA, J. M.; LOZANO, J. A.; LARRANAGA, P. An empirical comparison of four initialization methods for the k-means algorithm. **Pattern recognition letters**, Elsevier, v. 20, n. 10, p. 1027–1040, 1999. Citado na página 29.

PEREIRA, A. R.; ANGELOCCI, L. R.; SENTELHAS, P. C. **Meteorologia Agrícola**. Revista e ampliada. Piracicaba: ESALQ/USP - Departamento de Ciências Exatas, 2007. 1-3 p. Citado na página 16.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2025. Version 4.4.3. Disponível em: <https://www.R-project.org/>. Citado na página 24.

RAMALHO, M. A. P.; MARQUES, T. L.; LEMOS, R. d. C. Plant breeding in brazil: Retrospective of the past 50 years. **Crop Breeding and Applied Biotechnology**, SciELO Brasil, v. 21, p. e383021S3, 2021. Citado na página 20.

RAMSAY, J. O.; DALZELL, C. J. Some tools for functional data analysis. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Oxford University Press], v. 53, n. 3, p. 539–572, 1991. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2345586>. Citado na página 17.

SHARDA, R.; VOSS, S.; SUTHAHARAN, S. **Machine Learning Models and Algorithms for Big Data Classification; Thinking with Examples for Effective Learning**. [S.l.]: Springer: New York, NY, USA, 2019. Citado na página 30.

SIMPSON, Z. P.; HAGGARD, B. E. Optimizing the flow adjustment of constituent concentrations via loess for trend analysis. **Environmental Monitoring and Assessment**, v. 190, n. 2, p. 103, 2018. ISSN 1573-2959. Disponível em: <https://doi.org/10.1007/s10661-018-6461-5>. Citado na página 20.

SONG, H. *et al.* Hybrid causality analysis of enso's global impacts on climate variables based on data-driven analytics and climate model simulation. **Frontiers in Earth Science**, Volume 7 - 2019, 2019. ISSN 2296-6463. Disponível em: <https:

//www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2019.00233>. Citado na página 14.

ŠTRUMBELJ, E.; KONONENKO, I. Explaining prediction models and individual predictions with feature contributions. **Knowledge and information systems**, Springer, v. 41, n. 3, p. 647–665, 2014. Citado 4 vezes nas páginas 21, 22, 31 e 41.

TARPEY, T.; KINATEDER, K. K. Clustering functional data. **Journal of classification**, v. 20, n. 1, 2003. Citado 2 vezes nas páginas 18 e 25.

THOMPSON, J. R. Invited commentary: Re:'multiple comparisons and related issues in the interpretation of epidemiologic data". **American journal of epidemiology**, v. 147, n. 9, 1998. Citado na página 20.

ULLAH, S.; FINCH, C. F. Applications of functional data analysis: A systematic review. **BMC medical research methodology**, Springer, v. 13, n. 1, p. 43, 2013. Citado na página 15.

WALKER, G. T. Correlation in seasonal variations of weather—a further study of world weather. **Monthly Weather Review**, v. 53, n. 6, p. 252–254, 1925. Citado na página 16.

YASHODHA, G. *et al.* Efficient plant disease detection using k-means clustering and densenet-based classification. In: IEEE. **2025 International Conference on Electronics and Renewable Systems (ICEARS)**. [S.l.], 2025. p. 1197–1204. Citado na página 18.

YILDIRIM, C.; FRANCO-PEREIRA, A. M.; LILLO, R. E. Condition monitoring and multi-fault classification of hydraulic systems using multivariate functional data analysis. **Heliyon**, Elsevier, v. 11, n. 1, 2025. Citado na página 17.

ZHOU, P.; LIU, Z.; CHENG, L. An alternative approach for quantitatively estimating climate variability over china under the effects of enso events. **Atmospheric Research**, v. 238, p. 104897, 2020. ISSN 0169-8095. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169809519310609>. Citado na página 14.