Revista Pleiade – Centro Universitário Descomplica UniAmérica
ISSN 1980-8666 (Impresso) – ISSN 2674-8231 (Eletrônico) – Classificação Capes B3 (2017-2020)

Aplicação de Machine Learning na Previsão da Produtividade da Soja

Machine Learning Application in Soybean Productivity Forecasting

Eduardo Mendes Pereira[1], João Pedro dos Santos Beker[2] e Ruminiki Schmoeller[3]

1. Acadêmico concluinte do curso de Bacharelado em Engenharia de Software do Centro Universitário Descomplica UniAmérica. 2. Acadêmico concluinte do curso de Bacharelado em Engenharia de Software do Centro Universitário Descomplica UniAmérica. 3. Bacharel em Sistemas de Informação. Mestre em Tecnologias Computacionais para o Agronegócio. Docente do curso de Engenharia de Software do Centro Universitário Descomplica UniAmérica. https://orcid.org/0009-0006-5046-4390

eduardompereira4@gmail.com e *ruminiki.schmoeller@descomplica.com.br*

Palavras-chave
Aprendizado de máquina
Dados meteorológicos
Evapotranspiração
Previsão da produtividade
Soja

Keywords
Machine learning
Weather data
Evapotranspiration
Yield forecasting
Soybean

Resumo:
Este artigo apresenta estudo sobre a aplicação de técnicas de aprendizado de máquina para estimar a produtividade da soja na região oeste do Paraná, utilizando dados agrometeorológicos como temperatura, umidade relativa, radiação solar e evapotranspiração. O objetivo foi desenvolver modelos que auxiliem na previsão da produtividade, fornecendo informações valiosas para a gestão agrícola. A metodologia incluiu o uso de três modelos de aprendizado de máquina: Regressão Linear, Random Forest e Extreme Gradient Boosting, comparados em termos de acurácia e desempenho preditivo. O estudo se baseou em dados coletados entre 2008 e 2022 de 47 municípios do oeste do Paraná. Dentre os modelos, o Random Forest foi o que teve o melhor desempenho, com um coeficiente de determinação (R^2) de 0,86 no conjunto de treino e 0,81 no conjunto de teste. Em comparação, o Extreme Gradient Boosting apresentou R^2 de 0,77 no treino e 0,71 no teste, enquanto o modelo de Regressão Linear foi o de menor precisão, com R^2 de apenas 0,03. Conclui-se que os modelos de aprendizado de máquina são ferramentas com uma boa acurácia para otimizar a estimativa da produtividade de soja, possibilitando tomadas de decisão mais precisas e eficazes no campo.

Abstract:
This paper presents a study on the application of machine learning techniques to estimate soybean productivity in the western region of Paraná, using agrometeorological data such as temperature, relative humidity, solar radiation, and evapotranspiration. The objective was to develop models that assist in predicting productivity, providing valuable information for agricultural management. The methodology included the use of three machine learning models: Linear Regression, Random Forest, and Extreme Gradient Boosting, compared in terms of accuracy and predictive performance. The study was based on data collected between 2008 and 2022 from 47 municipalities in western Paraná. Among the models, Random Forest performed best, with a coefficient of determination (R^2) of 0.86 in the training set and 0.81 in the test set. In comparison, Extreme Gradient Boosting presented R^2 of 0.77 in training and 0.71 in testing, while the Linear Regression model had the lowest accuracy, with R^2 of only 0.03. It is concluded that machine learning models are tools with good accuracy to optimize the estimation of soybean productivity, enabling more precise and effective decision-making in the field.

Artigo recebido em: 16.10.2024.
Aprovado para publicação em: 14.11.2024.

1. INTRODUÇÃO

A soja (Glycine max L.) é uma das principais culturas agrícolas do mundo e desempenha um papel relevante na economia agrícola brasileira. O Brasil, maior produtor global de soja, registrou uma produção de 154,6 milhões de toneladas na safra de 2022/2023, com uma área plantada de aproximadamente 38,2 milhões de hectares (CONAB, 2020). No entanto, a safra 2023/2024 apresentou uma redução de 5,2%, totalizando 146,52 milhões de toneladas, principalmente devido a condições climáticas adversas, como baixas precipitações e temperaturas elevadas nas principais regiões produtoras (CONAB, 2024).

O estado do Paraná, o segundo maior produtor nacional de soja, sofreu uma redução de 17,9% na produtividade média durante a safra 2023/2024, resultando em 3.170 kg/ha (CONAB, 2024). As variações nas condições meteorológicas, como volume e distribuição das chuvas, são fatores críticos que afetam diretamente o desempenho da cultura, especialmente em fases fenológicas sensíveis, como o florescimento e o enchimento de grãos (INMET, 2009; FEHR; CAVINESS, 1977).

De acordo com Blanc e Shlenker (2017), a precipitação, temperatura e evapotranspiração são determinantes e afetam de forma significativa a produtividade das lavouras de soja. A evapotranspiração, em particular, é um componente essencial do balanço hídrico e energético, servindo como indicador fundamental para o manejo da irrigação e a otimização do rendimento agrícola (SILVA, 2018). O método de Penman-Monteith, recomendado pela FAO (*Food and Agriculture Organization*), é amplamente utilizado para estimar a evapotranspiração de referência (ET_0) a partir de dados meteorológicos, fornecendo maior precisão nas estimativas devido à inclusão de fatores aerodinâmicos e de resistência da superfície (ALLEN et al., 1998; OLIVEIRA, 2003).

A aquisição de dados meteorológicos com adequada resolução espaço-temporal e baixa taxa de falhas é um desafio recorrente em estudos agrometeorológicos (MARTINS et al., 2022). A plataforma NASA/POWER se destaca como uma ferramenta robusta para a obtenção de dados climáticos globais, fornecendo séries temporais detalhadas que podem ser utilizadas na modelagem de fenômenos agrícolas (GIOVANELLA et al., 2021).

O uso de técnicas de aprendizado de máquina (*Machine Learning*) tem se mostrado promissor para prever a produtividade agrícola, ao capturar relações complexas entre variáveis climáticas e de solo. Modelos como Redes Neurais, *Random Forest* e *Extreme Gradient Boosting* têm apresentado alta acurácia na predição da produtividade da soja em diferentes regiões (FERREIRA; CUNHA, 2020).

Dessa forma, o presente estudo tem como objetivo desenvolver modelos de *machine learning* para estimar a produtividade da soja no oeste do Paraná, utilizando dados de evapotranspiração e variáveis agrometeorológicas. A aplicação desses modelos visa fornecer previsões mais precisas e confiáveis, auxiliando na gestão agrícola e contribuindo para a sustentabilidade e competitividade da produção de soja na região.

2. ESTADO DA ARTE / TRABALHOS RELACIONADOS

2.1 MODELO DE ESTIMATIVA PARA A PREVISÃO METEOROLÓGICA PARA FINS AGRÍCOLAS UTILIZANDO *MACHINE LEARNING*

Vieira (2022) propôs um modelo de estimativa para a previsão meteorológica aplicada ao setor agrícola, utilizando técnicas de *Machine Learning*. O objetivo principal foi desenvolver um modelo de Regressão Li-

Pleiade, *18*(45): 25-38, Out.-Dez., 2024
DOI: 10.32915/pleiade.v18i45.1069

near Múltipla (RLM) para prever elementos meteorológicos com antecedência de dois meses em 15 localidades importantes na produção de milho no Brasil. O estudo utilizou dados diários de temperatura, umidade relativa, radiação solar e precipitação, provenientes da plataforma NASA/POWER e da Agência Nacional de Águas (ANA), organizados em períodos decendiais.

Os dados foram processados para prever variáveis como precipitação, temperatura média, mínima e máxima, velocidade do vento e outros elementos climáticos que afetam diretamente a produção agrícola, especialmente em climas do tipo Am e Aw. As previsões de precipitação apresentaram um R^2 ajustado acima de 0,62 em climas do tipo Am, com erros sistemáticos (ES) e raiz do erro quadrático médio (RMSE) satisfatórios. A RLM mostrou-se eficaz para a previsão em escala decendial, auxiliando no manejo de cultivos como o milho.

O estudo reforça a importância da previsão meteorológica para tomadas de decisão no campo, especialmente para reduzir os riscos climáticos e otimizar a produtividade agrícola.

2.2 Aprendizado de Máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo

O estudo de Guimarães (2019), concentra-se na aplicação de três modelos principais: Redes Neurais *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*. Esses modelos foram utilizados para prever a produtividade da soja em 27 cidades do estado do Mato Grosso, utilizando dados meteorológicos e de solo coletados entre 2010 e 2018. Posteriormente tiveram seus resultados comparados entre si com o modelo de estimativa de produtividade adotado pela FAO.

A pesquisa ressalta a importância de estimar o rendimento das culturas para melhorar a tomada de decisões no setor agrícola, permitindo otimização no manejo de culturas, controle de pragas, e ajuste de políticas públicas relacionadas à segurança alimentar. As previsões baseadas em técnicas de aprendizado de máquina são comparadas com modelos tradicionais, como o modelo agrometeorológico proposto por Doorenbos e Kassan (1979), que leva em conta a evapotranspiração e o balanço hídrico.

A análise demonstra que os modelos baseados em aprendizado de máquina, especialmente o *Extreme Gradient Boosting*, têm maior acurácia nas previsões da produtividade da soja, com base em variáveis climáticas e de solo, oferecendo assim uma ferramenta mais robusta e ajustada às condições reais observadas nas cidades estudadas. O estudo conclui que os três modelos tiveram ótima performance na predição da produtividade da soja.

2.3 Estimação da produtividade de soja a partir de modelo agrometeorológico com base em inteligência artificial

Santos (2023) investigou o potencial da inteligência artificial na previsão da produtividade da soja no Brasil, utilizando uma série temporal de 30 anos de dados climáticos (1988-2018) provenientes da NASA-POWER. O estudo empregou algoritmos de aprendizado de máquina, como *Random Forest* (RF), *Support Vector Machine* (SVM) e *Multilayer Perceptron* (MLP), para modelar a relação entre variáveis meteorológicas (precipitação, temperatura, velocidade do vento, umidade relativa e radiação solar) e a produtividade da soja em diferentes estádios fenológicos. Uma análise de correlação de Pearson foi realizada para identificar as variáveis meteorológicas mais influentes em cada fase do desenvolvimento da cultura.

Os resultados indicaram que os fatores climáticos exercem um impacto significativo e variável sobre a produtividade da soja, dependendo do estádio fenológico. O algoritmo de *Random Forest* apresentou o melhor desempenho, com um coeficiente de determinação ajustado (R^2) de 0,76 em algumas localidades, demonstrando sua capacidade de capturar as complexidades das relações entre as variáveis.

3. Materiais e Métodos

3.1 Seleção de Dados

O estudo foi realizado em 47 municípios da região oeste do Paraná. Os dados de produtividade foram obtidos por meio do banco de dados da Secretaria da Agricultura e do Abastecimento do Governo do Paraná (SEAB), referentes ao período de 2008 a 2022.

Para cada requisição de dados agrometeorológicos, foram informadas as coordenadas geográficas dos municípios, obtidas a partir do site do Instituto Brasileiro de Geografia e Estatística (IBGE). Os atributos coletados foram temperatura média (°C), radiação solar global (MJ/m²/dia), umidade relativa média (%) e velocidade do vento (m/s), referentes aos meses de setembro a janeiro, dos anos de 2008 a 2022, totalizando aproximadamente 2,5 milhões de registros (2.588.760). A definição do período foi feita com base no calendário de semeadura da soja no Paraná - que ocorre predominantemente entre setembro e dezembro, com a colheita concentrada nos meses de janeiro e fevereiro - e a disponibilidade de dados de produtividade anual (SEAB, 2024).

3.2 Pré-processamento

Nessa etapa foram realizados tratamentos de dados ausentes e *outliers*, substituindo dados faltantes pela mediana de cada atributo. Foram considerados *outliers*, os valores fora dos limites inferior e superior, definidos pela fórmula dada nas equações (1) e (2):

$$\textit{Limite Inferior:}\ Q1 - 1{,}5 \times IQR \quad (1)$$

$$\textit{Limite Superior:}\ Q3 + 1{,}5 \times IQR \quad (2)$$

onde: Q1 se refere ao primeiro quartil, Q3 ao terceiro quartil e IQR ao intervalo interquartil, ou seja, a diferença entre Q1 e Q3, conforme metodologia proposta por Nnamoko e Korkontzelos (2020). Na Tabela 1 pode-se observar as proporções de outliers detectados.

Tabela 1: Proporções de outliers

Coluna	Porcentagem de Outliers
WS2M	2,68
ALLSKY_SFC_SW_DWN	0,81
RH2M	0,1
T2M	0,7

Para o tratamento desses valores discrepantes, optou-se pela substituição dos *outliers* pela mediana, de modo a minimizar o impacto desses dados anômalos nas análises subsequentes.

3.3 TRANSFORMAÇÃO

As temperaturas máxima (°C) e mínima (°C) foram derivadas a partir dos valores extremos observados nas séries de temperatura média (°C) para cada localidade estudada. A evapotranspiração de referência (ETo) foi calculada para todas as localidades estudadas conforme o método de Penman-Monteith. Os dados meteorológicos diários usados para a obtenção da evapotranspiração de referência diária foram: temperatura máxima, temperatura mínima, temperatura média, radiação solar global, umidade relativa média e velocidade do vento.

$$ETo = \frac{0,408(R\eta - G) + \gamma \frac{900}{T+273} \mu_2 (e_s - e_a)}{\Delta + \gamma(1 + 0,3\mu_2)} \quad (3)$$

em que, **ETo** é a evapotranspiração de referência (mm d-1), **Rn** é o saldo de radiação à superfície da cultura (MJ m-2d-1), **G** é a densidade do fluxo de calor do solo (MJ m-2d-1), **T** é a temperatura do ar a 2 m de altura (°C), **u2** é a velocidade de vento a 2 m de altura (m s-1), **es** é a pressão de vapor de saturação (kPa), **ea** é a pressão parcial de vapor (kPa), **Δ** é a declividade da curva de pressão de vapor de saturação (kPa °C-1), e **γ** é o coeficiente psicrométrico (kPa °C-1).

Para que fosse possível a associação dos dados agrometeorológicos com os dados da produtividade da soja, foi necessário deixá-los na mesma granularidade de tempo. Enquanto os dados agrometeorológicos eram diários e estavam inicialmente em uma escala horária, os dados de produtividade eram anuais. Para alinhar essas escalas, foram transformados os dados agrometeorológicos de hora para dia, a partir da média diária. Em seguida, foram coletados os dados climáticos referentes aos meses de setembro, outubro, novembro, dezembro e janeiro de cada ano. Foi calculada a média de cada variável agrometeorológica nesse período. Esses valores médios foram então incorporados aos dados de produtividade da soja. Ao final desta etapa obteve-se um conjunto de dados com 10.906 observações.

3.4 MINERAÇÃO

Os modelos de Machine Learning utilizados foram desenvolvidos utilizando diferentes métodos e classes presentes na biblioteca Scikit-Learn[1], que é um projeto Python de código aberto com uma enorme comunidade presente, realizando contribuições constantemente. Em conjunto, foi utilizada a biblioteca Penmon[2] para calcular a evapotranspiração de referência ETo com base nos dados agrometeorológicos obtidos.

Para o modelo de *Linear Regression* (GALTON, 1886), os dados coletados foram separados em 80% para o conjunto de treino e 20% para o conjunto de teste. Também foi passado um parâmetro de controle para a geração de números aleatórios no modelo *random_state* com o valor 49. As variáveis independentes (X*1*… X*n*) utilizadas foram: evapotranspiração, umidade relativa média, velocidade do vento, temperatura máxima, temperatura mínima, temperatura média e radiação solar global. Como variável dependente Y foi utilizada a produtividade em kg/ha.

O modelo de *Random Forest* (BREIMAN, 2001) foi treinado para estimar a produtividade da colheita da soja utilizando as mesmas variáveis e conjunto de dados usados anteriormente no modelo de *Linear Regression* para poder realizar uma comparação entre os resultados obtidos.

O modelo *Extreme Gradient Boosting* (CHEN; GUESTRIN, 2016) foi treinado com as mesmas variáveis e conjunto de dados empregados no treinamento dos modelos de *Linear Regression* e *Random Forest*.

3.5 Interpretação dos Resultados

Para avaliar os diferentes modelos foram utilizadas duas métricas principais: coeficiente de determinação R^2 e a Raiz do Erro Quadrático Médio (RMSE). O coeficiente de determinação R^2 indica a proporção da variabilidade dos dados explicada pelo modelo, o que ajuda a avaliar sua capacidade de ajuste (SMITH, 2020). Além disso, o RMSE reflete uma métrica amplamente utilizada e reconhecida na comunidade de machine learning para medir o desempenho de modelos de regressão (FILHO, 2023).

Na etapa de interpretação foram analisados gráficos de comparação entre a produtividade real e a produtividade estimada de cada modelo.

4. Resultados e Discussão

4.1 Análise Exploratória dos Dados

Para entender melhor os dados utilizados neste estudo, foi realizada uma análise exploratória com foco nos dados meteorológicos e de produtividade da soja. As variáveis analisadas incluem temperatura média, radiação solar global, umidade relativa e velocidade do vento, todas elas importantes para estimar a produtividade da soja.

4.1.1 Correlação das variáveis utilizadas nos modelos

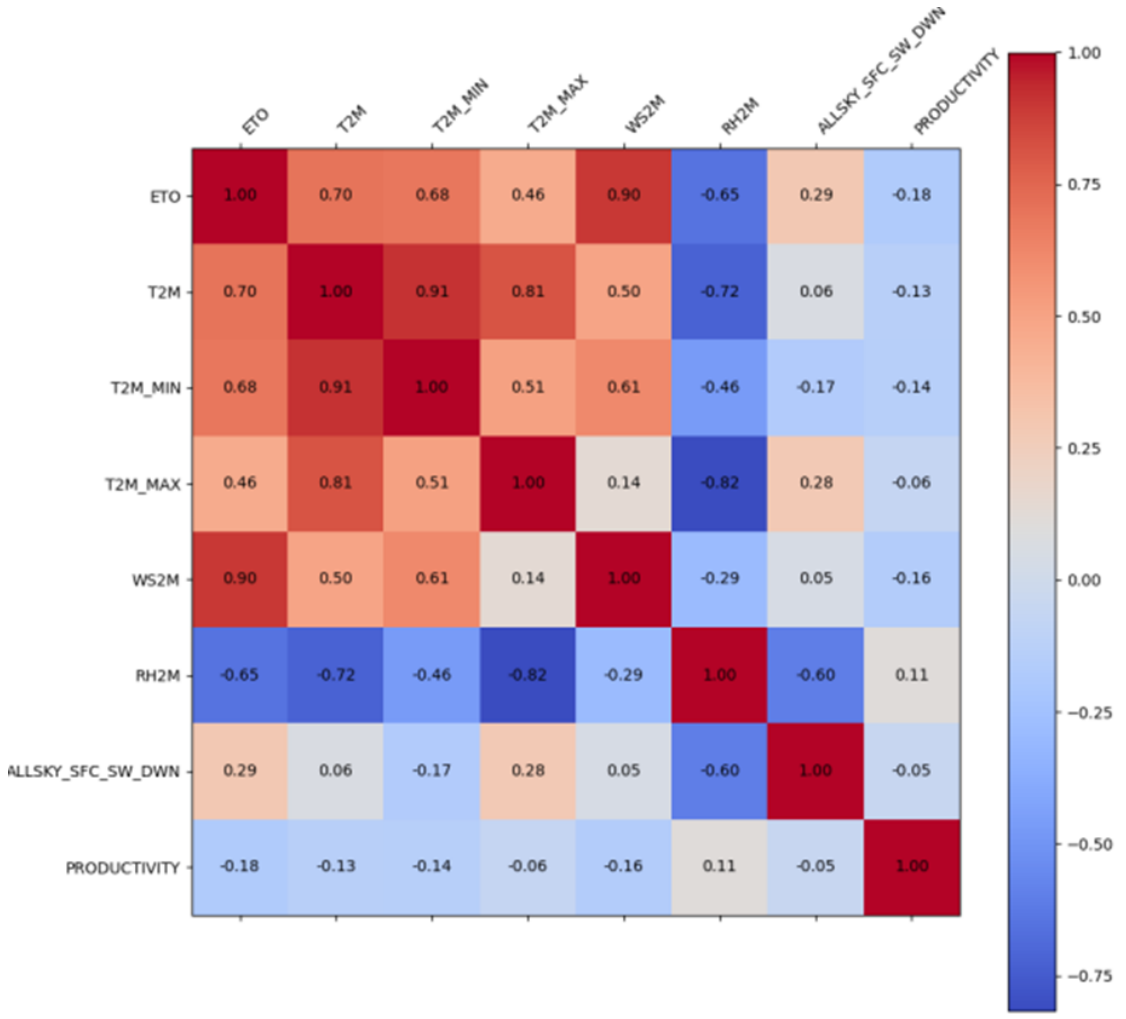
A Figura 1 exibe a correlação entre as variáveis climáticas e de produtividade da soja utilizadas nos diferentes modelos desenvolvidos.

É possível observar no gráfico que há uma baixa correlação entre a produtividade e as variáveis climáticas, justificando a demanda da estimativa da produtividade através do uso de algoritmos de *Machine Learning*.

4.1.2 Estatísticas Descritivas da Produtividade de Soja

Na Tabela 2, são apresentados os 20 municípios com maior produtividade do período de 2008 a 2022. As estatísticas incluídas são a média, mediana, desvio padrão, bem como os valores mínimo e máximo de produtividade (em kg/ha) ao longo dos anos analisados. Com base nos dados apresentados, é possível observar que os municípios de Corbélia, Iguatu e Boa Vista da Aparecida lideram com as maiores médias de produtividade, enquanto municípios como Ubiratã e São Pedro do Iguaçu apresentaram maiores variações ao longo do tempo, refletidas pelos altos desvios padrões.

Figura 1: Correlação entre as variáveis agrometeorológicas e de produtividade - Mapa de calor.



Fonte: Elaborada pelos autores.

Os resultados apresentados na Tabela 2 demonstram uma variabilidade na produtividade de soja entre os 20 municípios mais produtivos do período analisado. Embora Corbélia, Iguatu e Boa Vista da Aparecida tenham se destacado com as maiores médias de produtividade, a análise dos desvios padrões revela consideráveis oscilações nos rendimentos ao longo dos anos, particularmente em municípios como Ubiratã e São Pedro do Iguaçu. Essa heterogeneidade sugere a influência de diversos fatores, tanto climáticos quanto relacionados às práticas de manejo e características de cada localidade, os quais demandam investigações mais aprofundadas para uma compreensão completa dos padrões de produtividade da soja na região.

A Figura 2 evidencia a distribuição espacial da produtividade média de soja nos municípios do oeste do Paraná, com a legenda indicando a classificação em três categorias: baixa (até 2.900 kg/ha), média (entre 2.901 e 3.300 kg/ha) e alta (acima de 3.300 kg/ha).

Tabela 2: Estatísticas Descritivas da Produtividade de Soja nos 20 Municípios com maior Produtividade (2008-2022)

Município	Média	Mediana	Desvio Padrão	Min	Max
Corbélia	3.640,13	3716	481,72	2515	4250
Iguatu	3.534,87	3600	355,50	2611	4003
Boa Vista da Aparecida	3.490,67	3571	503,56	2381	4452
Santa Tereza do Oeste	3.485,13	3450	410,76	2739	4350
Cafelândia	3.454,73	3460	714,41	1450	4515
Santa Lúcia	3.437,47	3422	325,94	2913	4150
Capitão Leônidas Marques	3.420,13	3401	386,18	2432	4200
Braganey	3.398,93	3390	368,57	2708	3850
Cascavel	3.381,27	3464	422,71	2549	4119
Anahy	3.350,73	3471	526,39	1893	4091
Céu Azul	3.339,40	3591	746,08	1734	4100
Formosa do Oeste	3.338,27	3545	760,83	1600	4300
Lindoeste	3.298,07	3250	326,24	2661	3900
Jesuítas	3.219,80	3458	911,92	917	4200
Tupãssi	3.214,27	3451	834,00	1165	4200
Nova Aurora	3.198,53	3459	867,11	1200	4282
Santa Terezinha de Itaipu	3.187,20	3400	819,67	700	4200
Vera Cruz do Oeste	3.174,13	3470	843,91	1305	4091
Ubiratã	3.124,13	3343	653,34	1200	4090
São Pedro do Iguaçu	3.097,80	3393	754,33	1248	4000

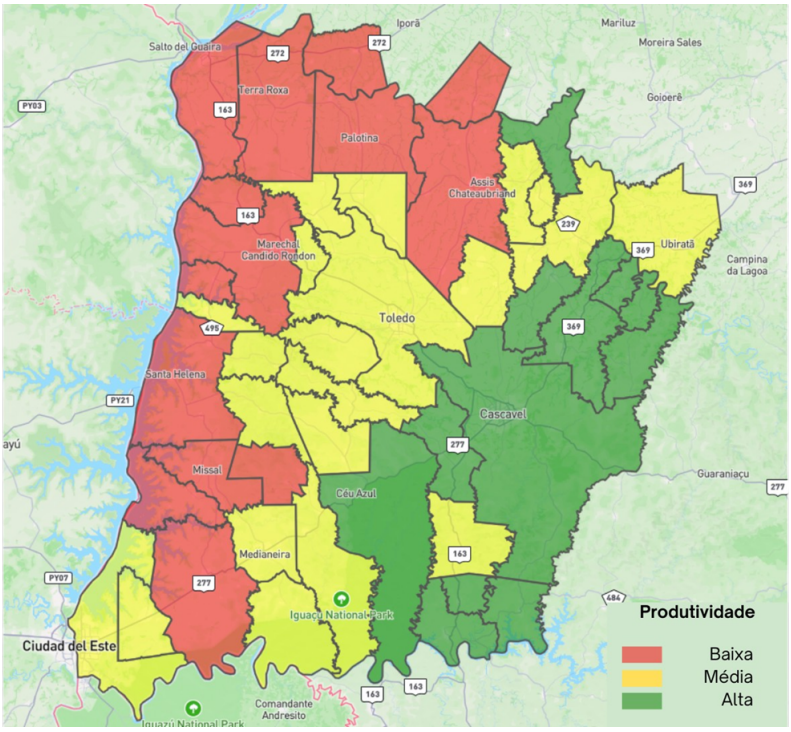
A visualização do mapa permite identificar regiões com maior potencial produtivo, bem como áreas que demandam maior atenção em termos de manejo e investimento. A concentração de áreas de alta produtividade em determinadas regiões pode estar associada a fatores como condições climáticas mais favoráveis, adoção de tecnologias mais avançadas e maior acesso a insumos e serviços.

4.1.3 MÉDIA DAS VARIÁVEIS AGROMETEOROLÓGICAS

A Tabela 3 a seguir, apresenta a média e o desvio padrão dos dados climáticos coletados anualmente de 2008 a 2022, incluindo variáveis como temperatura média (T2M), temperatura máxima (T2M_MAX), temperatura mínima (T2M_MIN), radiação solar global incidente (ALLSKY_SFC_SW_DWN), umidade relativa do ar (RH2M), velocidade do vento (WS2M) e evapotranspiração potencial (ETO). Esses dados são essenciais para compreender as variações climáticas ao longo dos anos e suas implicações em diversas áreas, como agricultura, gestão de recursos hídricos e previsão de fenômenos meteorológicos.

As variáveis agrometeorológicas apresentadas na Tabela 3, coletadas ao longo de 15 anos (2008-2022), revelam alguns dados interessantes. A temperatura média (T2M) foi de 24,05 °C, enquanto as máximas (T2M_MAX) e mínimas (T2M_MIN) ficaram em 29,59 °C e 18,99 °C, respectivamente. Isso indica uma amplitude térmica significativa, que pode impactar o crescimento das culturas na região.

Figura 2: Mapa da média da produtividade de soja nos municípios do oeste do Paraná.



Fonte: Elaborada pelos autores.

Tabela 3: Média das variáveis agrometeorológicas (2008-2022)

	T2M	T2M_MAX	T2M_MIN	ALLSKY_SFC_SW_DWN	RH2M	WS2M	ETO
Média	24,05	29,59	18,99	0,86	73,67	0,99	1,21
Desvio Padrão	1,38	1,51	1,63	0,04	5,29	0,57	0,38

A radiação solar global (ALLSKY_SFC_SW_DWN) teve uma média de 18,06 MJ/m²/dia, um fator crucial para evapotranspiração e produtividade agrícola. A umidade relativa do ar (RH2M) registrou uma média de 73,67%, proporcionando condições de moderada a alta umidade, favorável na cultura da soja. A velocidade do vento (WS2M) média foi de 0,98 m/s, característica de áreas continentais com baixa circulação de ar.

A evapotranspiração potencial (ETO) apresentou uma média de 4,21 mm/dia. O desvio padrão das variáveis demonstra a variabilidade climática, indicando estabilidade ou oscilação ao longo dos anos.

4.2 Desempenho dos Modelos

Ao comparar os modelos *Linear Regression*, *Random Forest* e *Extreme Gradient Boosting*, observou-se que o modelo de *Random Forest* apresentou o melhor desempenho tanto no conjunto de treino quanto no conjunto de teste.

Para o conjunto de treino, o modelo *Random Forest* obteve um coeficiente de determinação (R^2) de 0,86, indicando que 86% da variabilidade da produtividade pode ser explicada pelas variáveis independentes. O erro quadrático médio (RMSE) foi de 309,33, sugerindo um desvio médio de 309,33 kg/ha. Em comparação, o

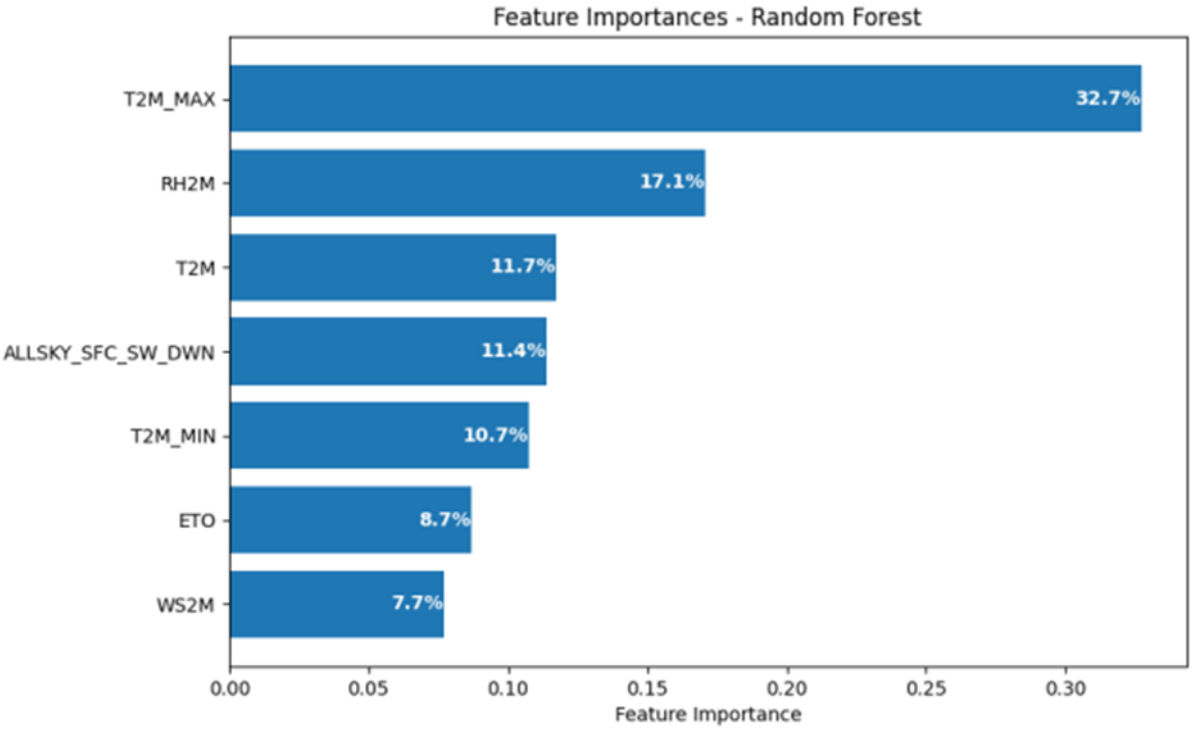
modelo *Extreme Gradient Boosting* obteve um R^2 de 0,77 e um RMSE de 401,67. O modelo *Linear Regression*, por sua vez, teve um desempenho inferior, com um R^2 de 0,03 e um RMSE de 823,82.

No conjunto de teste, o *Random Forest* continuou destacando-se com um R^2 de 0,81 e um RMSE de 410,18. O *Extreme Gradient Boosting* apresentou um R^2 de 0,71 e um RMSE de 507,19. Já o modelo de *Linear Regression* teve um desempenho significativamente inferior, com um R^2 de 0,02 e um RMSE de 934,09.

A baixa precisão da Regressão Linear pode ser explicada pela sua incapacidade de capturar a complexidade das interações entre variáveis climáticas e de produtividade, especialmente em cenários onde a correlação entre essas variáveis é baixa. Esse comportamento também foi observado no estudo de Tatiana da Silva (2023), no qual a Regressão Linear Múltipla (RLM) apresentou um R^2 ajustado de apenas 0,01 no grupo climático 4, caracterizado por alta variabilidade meteorológica e complexidade climática. Esses achados indicam que a Regressão Linear, embora útil em contextos mais simples, não é adequada para cenários agrometeorológicos complexos e heterogêneos.

A Figura 3 a seguir exibe as variáveis com maior relevância no modelo de Random Forest implementado para análise preditiva. As barras no gráfico representam a importância relativa de cada variável, destacando aquelas que mais influenciam as previsões do modelo.

Figura 3: Variáveis de maior importância no modelo Random Forest



Fonte: Elaborada pelos autores.

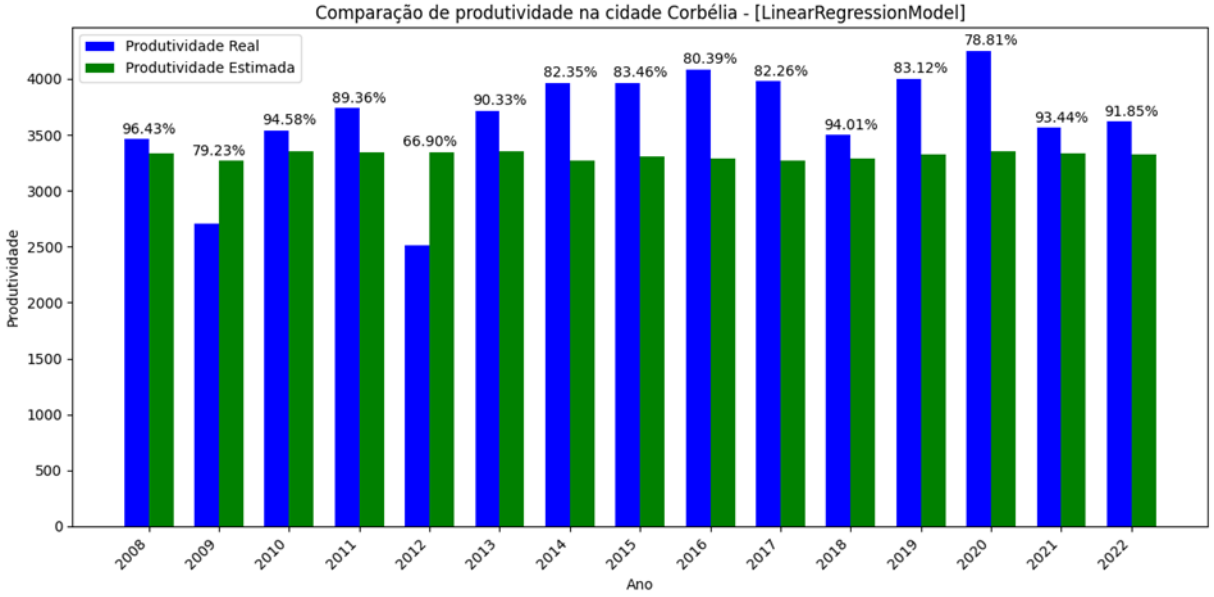
Conforme apresentado na Figura 3, observa-se que a variável T2M_MAX (temperatura máxima ao nível de 2 metros) é a mais significativa, representando 32,7% da importância total no modelo. Em seguida, a umidade relativa ao nível de 2 metros (RH2M) e a temperatura média ao nível de 2 metros (T2M) correspondem a 17,1% e 11,7%, respectivamente, reforçando a relevância dos parâmetros térmicos e de umidade para a previsão em análise. A variável de radiação solar (ALLSKY_SFC_SW_DWN) também apresentou uma contri-

buição notável, com 11,4%, evidenciando a influência da radiação no desempenho do modelo. As demais variáveis, incluindo a temperatura mínima (T2M_MIN), a evapotranspiração (ETO) e a velocidade do vento ao nível de 2 metros (WS2M), apresentam percentuais inferiores, mas, ainda assim, desempenham papel relevante na composição do modelo.

O estudo de Edson da Silva Guimarães (2019) corroborou a superioridade dos modelos de aprendizado de máquina em predições agrícolas, destacando o *Extreme Gradient Boosting* como o mais preciso, com 95,54% de acurácia. O modelo *Random Forest* também obteve um desempenho semelhante, com 95,03% de precisão, comprovando sua eficácia em modelar a produtividade agrícola em condições variadas. A semelhança entre os resultados do presente estudo e os de Guimarães evidencia que, em regiões agrícolas com alta heterogeneidade, tanto o *Random Forest* quanto o *Extreme Gradient Boosting* são modelos mais eficazes para capturar as interações complexas entre variáveis climáticas e de solo, fornecendo previsões mais precisas do que a Regressão Linear em cenários com alta variabilidade.

A Figura 4 mostra o desempenho obtido pelo modelo *linear regression* na estimativa da produtividade de soja entre os anos de 2008 a 2022 na cidade de Corbélia, através de um gráfico comparativo da produtividade real, representado pela barra de cor azul e a produtividade estimada pelo modelo, representada pela barra de cor verde, onde é possível observar que o modelo não conseguiu acompanhar a tendência da variabilidade da produtividade ao decorrer dos anos.

Figura 4: Acurácia do modelo *linear regression*

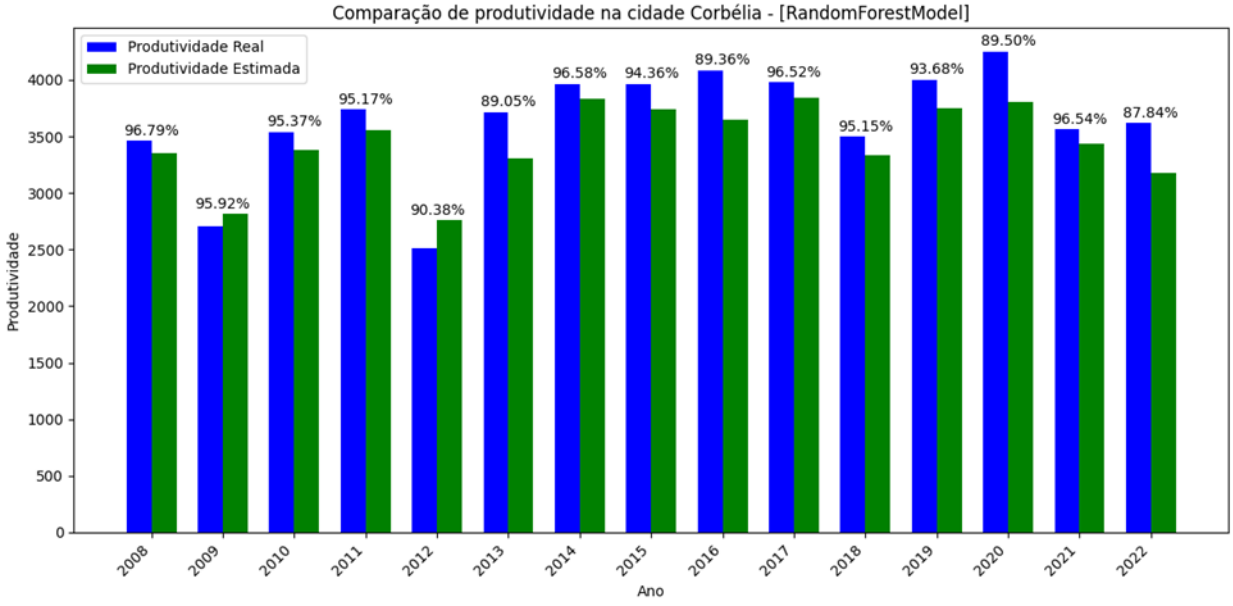


Fonte: Elaborada pelos autores.

Na Figura 5 é exibido o resultado de estimativa da produtividade de soja atingida pelo modelo *random forest*, utilizando os mesmos parâmetros comparativos do modelo *linear regression*. A estimativa obtida no modelo acompanha a tendência da produtividade real ao longo dos anos.

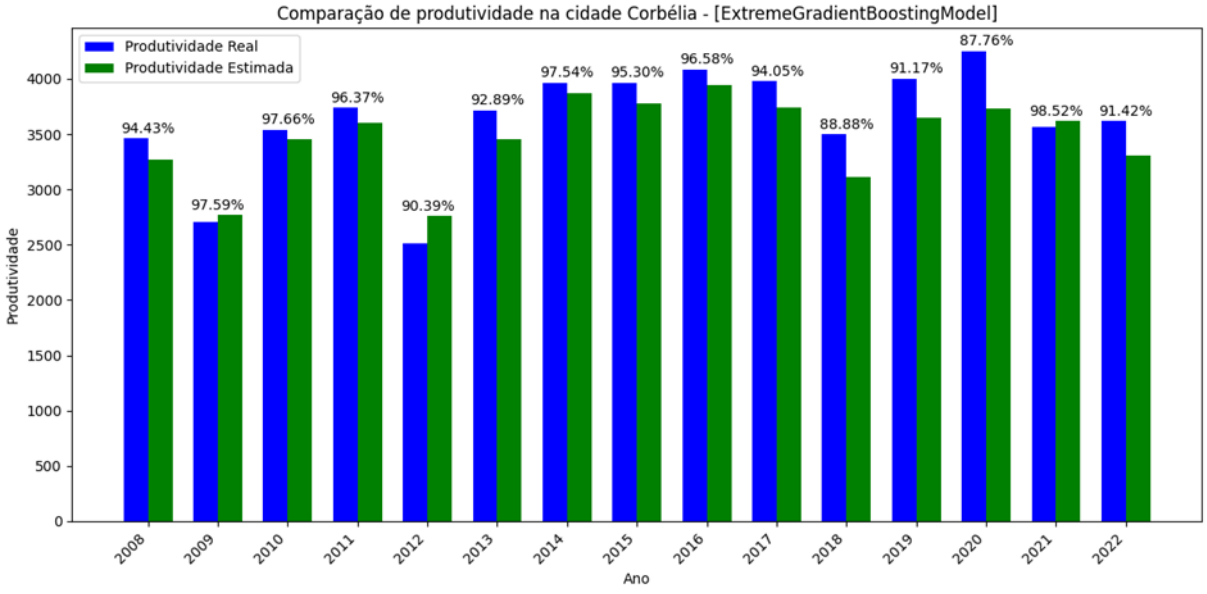
Já a Figura 6 apresenta a estimativa da produtividade de soja alcançada pelo modelo *extreme gradient boosting*, utilizando os mesmos parâmetros de comparação usados nos outros modelos.

Figura 5: Acurácia do modelo *random forest*



Fonte: Elaborada pelos autores.

Figura 6: Acurácia do modelo *extreme gradient boosting*



Fonte: Elaborada pelos autores.

Através da comparação com os valores reais de produtividade, representados visualmente na Figura 6, é possível observar que o Extreme Gradient Boosting conseguiu capturar as variações sazonais da produtividade, acompanhando as flutuações e tendências apresentadas nos dados históricos.

CONSIDERAÇÕES FINAIS

Este estudo demonstrou que os modelos de *Machine Learning* podem ser uma ferramenta útil para estimar a produtividade da soja no oeste do Paraná. Os resultados obtidos indicam que as variáveis climáticas, especialmente radiação solar global, umidade relativa média, velocidade do vento e evapotranspiração desempenham um papel importante na determinação da produtividade.

As estimativas de produtividade geradas por esses modelos podem auxiliar os agricultores a tomar decisões mais informadas sobre o manejo de suas lavouras, contribuindo para aumentar a eficiência e a sustentabilidade da produção de soja.

Para trabalhos futuros, sugere-se incluir outras variáveis relevantes, como a qualidade do solo, níveis de fertilização e histórico de pragas e doenças, a fim de aprimorar a precisão das previsões. Além disso, o uso de outras técnicas avançadas de *Machine Learning*, como redes neurais profundas, pode oferecer novos *insights* e melhorar a capacidade preditiva dos modelos. Também é recomendado avaliar o desempenho desses modelos em outras regiões produtoras de soja, para testar aplicabilidade e robustez em diferentes contextos.

NOTAS

1. SCIKIT-LEARN. Machine Learning in Python. Versão 1.5.1. Disponível em: <https://scikit-learn.org/\>. Acesso em: 26 out. 2024.
2. PENMON. Calculadora de Evapotranspiração. Versão 1.5. Disponível em: <https://github.com/sherzodr/penmon/\>. Acesso em: 26 out. 2024.

REFERÊNCIAS

ALLEN, R. G. et al. **Crop evapotranspiration: guidelines for computing crop water requirements**. Roma: FAO, 1998. 300 p.

BLANC, E.; SCHLENKER, W. The use of panel models in assessments of climate impacts on agriculture. *Review of Environmental Economics and Policy*, v. 11, n. 2, p. 258-279, 2017.

BREIMAN, L. **Random Forests**. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001. DOI: 10.1023/A:1010933404324.

CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016. p. 785-794. DOI: 10.1145/2939672.2939785.

CONAB. **Produção de grãos da safra 2020/21 segue como maior da história: 268,9 milhões de toneladas**. Brasília: Conab, 2020. Disponível em: https://www.conab.gov.br/ultimas-noticias/3691-producao-de-graos-da-safra-2020-21-segue-como-maior-da-historia-268-9-milhoes-de-toneladas. Acesso em: 21 abr. 2024.

CONAB. **Safra de grãos 2023/2024 está estimada em 294,1 milhões de toneladas**. Brasília: Conab, 2024. Disponível em: https://www.conab.gov.br/ultimas-noticias/5478-safra-de-graos-2023-2024-esta-estimada-em-294-1-milhoes-de-toneladas. Acesso em: 21 abr. 2024.

FEHR, W. R.; CAVINESS, C. E. **Stages of soybean development**. Ames: Iowa State University, Dept. of Science and Technology, 1977. 11 p. (Special report, 80).

FERREIRA, L. B.; CUNHA, F. **New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning**. *Agricultural Water Management*, v. 234, p. 106-113, 2020. Disponível em: https://doi.org/10.1016/j.agwat.2020.106113. Acesso em: 04 maio 2024.

FILHO, M. **RMSE (Raiz do erro quadrático médio) em machine learning**. 2023. Disponível em: https://mariofilho.com/rmse-raiz-do-erro-quadratico-medio-em-machine-learning. Acesso em: 22 jun. 2024.

GALTON, F. **Regression towards mediocrity in hereditary stature**. *The Journal of the Anthropological Institute of Great Britain and Ireland*, v. 15, p. 246-263, 1886.

GIOVANELLA, T. H.; OLIVEIRA, F. C.; MARCHI, V. A.; TLUSZCZ, J. **Desempenho de métodos de preenchimento de falhas em dados de evapotranspiração de referência para região oeste do Paraná**. *Revista Brasileira de Meteorologia*, v. 36, n. 3, p. 415-422, 2021.

GUIMARÃES, E. S. **Aprendizado de máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo.** Dissertação (Mestrado em Matemática, Estatística e Computação Aplicadas à Indústria) - Universidade de São Paulo, São Carlos, 2019.

INMET. **Agrometeorologia dos Cultivos: o fator meteorológico na produção agrícola**. Brasília: INMET, 2009.

MARTINS, L. L. et al. **Utilização dos dados do NASA-POWER em estudos agrometeorológicos: análise qualitativa da evapotranspiração de referência**. São Paulo: IAC, 2022.

NNAMOKO, N.; KORKONTZELOS, I. **Efficient treatment of outliers and class imbalance for diabetes prediction**. *Artificial Intelligence in Medicine*, v. 104, p. 101815, 2020.

OLIVEIRA, A. D. de. **Comparação de métodos de estimativa da evapotranspiração de referência utilizando dados de estação meteorológica convencional e automática**. 2003. 70 f. Tese (Doutorado em Produção Vegetal) - Universidade Estadual Paulista, Jaboticabal, 2003.

SANTOS, T. S. **Estimação da produtividade de soja a partir de modelo agrometeorológico com base em inteligência artificial**. 2023. Tese (Doutorado em Agronomia – Produção Vegetal) – Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, 2023.

SEAB. **Regionalização da semeadura de soja no Paraná: safra 2024/2025**. Paraná: Secretaria da Agricultura e do Abastecimento, 2024. Disponível em: http://agricultura.pr.gov.br. Acesso em: 11 maio 2024.

SILVA, Y. F. **Uso do algoritmo SAFER para evapotranspiração real na cultura da soja**. São Paulo: UNESP, 2018.

SMITH, J. **Understanding the Coefficient of Determination (R^2) in Regression Analysis**. *Journal of Statistical Methods*, v. 15, n. 2, p. 123-134, 2020.