DÁRIO DA SILVA MELO

GUIAFERTIL: SOFTWARE DE ANÁLISE EXPLORATÓRIA E PREDIÇÃO DO LIMIAR DE PRODUTIVIDADE DA CULTURA DE SOJA COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

MONTE CARMELO - MG
2025

DÁRIO DA SILVA MELO

GUIAFERTIL: SOFTWARE DE ANÁLISE EXPLORATÓRIA E PREDIÇÃO DO LIMIAR DE PRODUTIVIDADE DA CULTURA DE SOJA COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Dissertação apresentada ao Curso de Mestrado Acadêmico em Agricultura e Informações Geoespaciais da Universidade Federal de Uberlândia como requisito parcial para aprovação no curso de Mestrado.

Orientador: Dr. Murillo Guimarães Carneiro
Coorientadores: Dr. Odair José Marques e Dr. Douglas José Marques

MONTE CARMELO - MG
2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

M528 Melo, Dário da Silva, 1977-
2026 GUIAFERTIL: SOFTWARE DE ANÁLISE EXPLORATÓRIA E PREDIÇÃO DO LIMIAR DE PRODUTIVIDADE DA CULTURA DE SOJA COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL [recurso eletrônico] / Dário da Silva Melo. - 2026.

Orientador: Murillo Guimarães Carneiro .
Coorientador: Odair José Marques .
Coorientador: Douglas José Marques .
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Agricultura e Informações Geoespaciais.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2026.195
Inclui bibliografia.

1. Agronomia. I. , Murillo Guimarães Carneiro,1988-, (Orient.). II. , Odair José Marques,1973-, (Coorient.). III. , Douglas José Marques, 1980-, (Coorient.). IV. Universidade Federal de Uberlândia. Pós-graduação em Agricultura e Informações Geoespaciais. V. Título.

CDU: 631

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Agricultura e Informações Geoespaciais
Rodovia LMG 746, Km 01, s/nº, Bloco 1AMC, Sala 1A202 , Monte Carmelo-MG, CEP 38.500-000
Telefone: (34) 3810-1033 - ppgaig@iciag.ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Agricultura e Informações Geoespaciais				
Defesa de:	Dissertação de Mestrado Acadêmico				
Data:	20/01/2026	Hora de início:	8:00h	Hora de encerramento:	10:16h
Matrícula do Discente:	32222AIG002				
Nome do Discente:	Dário da Silva Melo				
Título do Trabalho:	GUIAFERTIL: SOFTWARE DE ANÁLISE EXPLORATÓRIA E PREDIÇÃO DO LIMIAR DE PRODUTIVIDADE NA CULTURA DE SOJA COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL				
Área de concentração:	Informações geoespaciais e tecnologias aplicadas à produção agrícola				
Linha de pesquisa:	Desenvolvimento e aplicações de métodos em informações geoespaciais				
Projeto de Pesquisa de vinculação:	Inteligência artificial aplicada à agricultura e informações geoespaciais				

Reuniu-se em "sala virtual" a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Agricultura e Informações Geoespaciais, composta pelo prof. Dr. Murillo Guimarães Carneiro (Universidade Federal de Uberlândia-UFU), orientador do candidato, prof. Dr. Daniel Stefany Duarte Caetano (Universidade Federal de Uberlândia-UFU), e prof. Dr. Hudson Carvalho Bianchini (UNIFENAS).

Iniciando os trabalhos o presidente da mesa, Dr. Murillo Guimarães Carneiro, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao discente a palavra para a exposição do seu trabalho. A duração da apresentação da discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Daniel Stefany Duarte Caetano**, **Professor(a) do Magistério Superior**, em 20/01/2026, às 10:18, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Murillo Guimarães Carneiro**, **Professor(a) do Magistério Superior**, em 20/01/2026, às 10:19, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Hudson Carvalho Bianchini**, **Usuário Externo**, em 20/01/2026, às 15:20, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6998057** e o código CRC **927E72CF**.

Referência: Processo nº 23117.002753/2026-95 SEI nº 6998057

BIOGRAFIA

Nascido em 13 de março de 1977, no município de Monte Carmelo, Minas Gerais, Dário da Silva Melo construiu uma carreira sólida marcada pela dedicação ao setor agropecuário, pela competência em gestão e tecnologia, e pela busca constante por inovação. Desde cedo, demonstrou grande interesse pelas áreas de tecnologia da informação e gestão empresarial, o que o levou a concluir duas graduações: Ciências Contábeis e Ciências da Computação pelo Centro Universitário do Triângulo (UNITRI). Essa combinação de formações lhe proporcionou uma visão ampla e estratégica, unindo conhecimento técnico e administrativo. Entre 2003 e 2013, atuou como analista de sistemas e programador na empresa Retiro da Roça, localizada em Monte Carmelo-MG. Nesse período, desenvolveu sistemas e soluções tecnológicas que otimizaram processos e reforçaram a eficiência operacional da empresa, contribuindo significativamente para sua modernização. A partir de 2014, assumiu a função de gerente administrativo na mesma empresa, posição que ocupa até hoje. Ao longo dessa trajetória, consolidou-se como um líder versátil e dedicado, responsável pela coordenação de compras, planejamento de campanhas agrícolas de soja e milho, gestão de barter (troca de insumos por grãos), além da administração de equipes de agrônomos de campo e da área de tecnologia da informação. Sua experiência abrange ainda o gerenciamento de atividades em fábrica de ração, armazéns gerais, revenda agrícola e pecuária, reforçando sua ampla visão do agronegócio e sua capacidade de integrar setores estratégicos. Além da prática profissional, Dário destaca-se pela habilidade didática no treinamento e capacitação de equipes, tanto administrativas quanto de produção, transmitindo conhecimento técnico de forma clara e acessível. Em 2022, essa vocação educadora resultou em seu ingresso no Programa de Mestrado em Agricultura e Informações Geoespaciais da Universidade Federal de Uberlândia.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

GLOSSÁRIO

Agricultura de precisão: Conjunto de práticas que utilizam tecnologias para coletar, processar e analisar dados sobre o solo, clima e planta, visando otimizar o manejo agrícola com sustentabilidade e eficiência.

Altitude: Distância vertical de um ponto em relação ao nível do mar; variável geoespacial que influencia o clima, a temperatura e a produtividade das culturas.

Análise exploratória de dados: Processo inicial de investigação de um conjunto de dados com o objetivo de resumir suas principais características, frequentemente com métodos visuais e estatísticos.

Aprendizado de máquina: Campo da inteligência artificial que permite que sistemas aprendam automaticamente com dados e realizem previsões sem serem explicitamente programados.

Argila: Partícula mineral fina do solo com alta capacidade de retenção de água e nutrientes, influenciando diretamente a fertilidade e estrutura do solo.

Base de dados: Conjunto estruturado de informações armazenadas digitalmente, utilizado para treinar, testar e validar modelos preditivos.

Batch size: Número de amostras processadas simultaneamente antes da atualização dos pesos durante o treinamento de uma rede neural.

Bias (viés): Parâmetro de ajuste de um neurônio artificial que desloca a função de ativação, permitindo maior flexibilidade no aprendizado.

CTC (Capacidade de Troca Catiônica): Propriedade do solo que indica sua capacidade de reter e fornecer nutrientes às plantas.

Camada de entrada: Primeira camada de uma rede neural, responsável por receber os dados brutos de entrada.

Camada de saída: Última camada de uma rede neural, responsável por fornecer o resultado ou a previsão final.

Camada oculta: Camadas intermediárias de uma rede neural, onde ocorre o processamento e abstração dos dados.

Cases Campeões (CESB): Base de dados de lavouras auditadas de alta produtividade da soja, organizada pelo Comitê Estratégico Soja Brasil.

Coeficiente de determinação (R^2): Métrica estatística que mede o quanto as variações de uma variável dependente podem ser explicadas pelo modelo.

Correlação: Medida estatística que expressa o grau de associação entre duas variáveis.

Dados categóricos: Variáveis qualitativas que representam categorias ou grupos sem relação numérica direta.

Dados quantitativos: Variáveis numéricas que expressam medições contínuas ou discretas.

Dataset: Conjunto de dados utilizado para treinar e avaliar modelos de aprendizado de máquina.

Epoch: Número de vezes que o conjunto de dados completo é passado ao modelo durante o processo de treinamento.

Épsilon (ε): Parâmetro da regressão por vetores de suporte (SVR) que define uma margem de tolerância dentro da qual os erros não são penalizados.

Erro absoluto médio (MAE): Métrica que calcula a média dos valores absolutos das diferenças entre as previsões e os valores reais.

Erro quadrático médio (RMSE): Métrica que calcula a raiz da média dos quadrados das diferenças entre as previsões e os valores reais.

Feature (atributo): Característica ou variável utilizada como entrada para um algoritmo de aprendizado de máquina.

Fertilizante: Produto químico ou orgânico aplicado ao solo para suprir nutrientes essenciais ao desenvolvimento das plantas.

Floresta Aleatória: Algoritmo de aprendizado de máquina baseado em múltiplas árvores de decisão, usado para melhorar a precisão das previsões por meio do método ensemble.

Função ReLU: Função de ativação que retorna o valor de entrada quando positivo e zero caso contrário; amplamente usada em redes neurais profundas.

Função de ativação: Função matemática aplicada à saída de um neurônio artificial para introduzir não linearidade ao modelo.

GUI (Interface Gráfica do Usuário): Componente visual de um software que permite a interação com o usuário por meio de botões, janelas e menus.

Guiafertil: Software desenvolvido nesta dissertação que integra análise exploratória e algoritmos de aprendizado de máquina para prever a produtividade da soja.

Hiperparâmetro: Parâmetro configurado antes do treinamento do modelo, como taxa de aprendizado, número de camadas ou árvores, que influencia o desempenho do modelo.

Inteligência Artificial (IA): Área da computação que desenvolve sistemas capazes de executar tarefas que requerem inteligência humana, como aprendizado e tomada de decisão.

Keras: Biblioteca de código aberto em Python usada para criar e treinar redes neurais, baseada no TensorFlow.

Kernel: Função matemática usada em SVM e SVR para transformar os dados em um espaço de maior dimensão, permitindo separações não lineares.

Learning rate (taxa de aprendizado): Parâmetro que controla a velocidade com que o modelo ajusta seus pesos durante o treinamento.

Limiar de produtividade: Valor estimado que representa o limite máximo de produtividade atingível com base nas condições de solo, clima e manejo.

Machine Learning supervisionado: Tipo de aprendizado de máquina em que o modelo é treinado com exemplos de entrada e saída conhecidos.

Margem de tolerância: Intervalo de erro permitido em torno da previsão no modelo SVR sem penalização.

Matriz de dados: Representação tabular dos atributos e instâncias usadas para treinar e testar modelos.

Modelo preditivo: Modelo matemático ou computacional capaz de estimar valores futuros com base em dados históricos.

Neurônio artificial: Unidade computacional que compõe as redes neurais, responsável por processar entradas ponderadas e gerar saídas.

Normalização: Processo de padronização de variáveis numéricas para reduzir distorções e melhorar o desempenho dos modelos.

Overfitting: Fenômeno em que o modelo aprende demais os dados de treinamento e perde capacidade de generalização para novos dados.

Perceptron Multicamadas (MLP): Tipo de rede neural com múltiplas camadas de neurônios que realiza aprendizado supervisionado.

pH: Medida que expressa o grau de acidez ou alcalinidade de uma solução; essencial para avaliação da fertilidade do solo.

Predição: Estimativa gerada por um modelo de aprendizado de máquina com base nos dados de entrada.

Python: Linguagem de programação utilizada para o desenvolvimento do software Guiafertil.

Reamostragem (Bootstrap): Técnica estatística usada para criar subconjuntos de dados de treinamento com reposição, comum em métodos ensemble.

Rede Neural: Modelo computacional inspirado no funcionamento do cérebro humano, composto por camadas de neurônios interconectados.

Regressão Linear: Método estatístico usado para modelar a relação entre uma variável dependente e uma ou mais independentes.

Reprodutibilidade: Capacidade de reproduzir resultados idênticos sob as mesmas condições experimentais, essencial em pesquisas científicas.

R^2: Símbolo usado para o coeficiente de determinação; mede a qualidade do ajuste de um modelo preditivo.

SQLite3: Sistema de banco de dados relacional leve, utilizado para armazenamento local de dados no software Guiafertil.

SVR (Regressão de Vetores de Suporte): Técnica de aprendizado de máquina usada para regressão, derivada do algoritmo SVM.

Seed (semente): Número inicial usado para gerar resultados aleatórios reprodutíveis em experimentos de aprendizado de máquina.

Sistema de Plantio Direto: Sistema agrícola que mantém a cobertura vegetal no solo, reduzindo erosão e preservando umidade.

Taxa de aprendizado: Ver Learning rate.

TensorFlow: Biblioteca de código aberto voltada ao aprendizado profundo e cálculos numéricos, base do Keras.

Treinamento: Etapa do aprendizado de máquina em que o modelo ajusta seus parâmetros internos com base nos dados de entrada e saída.

Validação cruzada: Método de avaliação que divide os dados em subconjuntos para testar o desempenho e evitar overfitting.

VANT (Veículo Aéreo Não Tripulado): Aeronave sem piloto embarcado utilizada na agricultura para monitoramento das lavouras por sensores embarcados e para aplicações aéreas, como pulverização e distribuição de insumos.

Variável dependente: Variável cujo valor é previsto pelo modelo, neste caso a produtividade da soja.

Variável independente: Atributos de entrada que influenciam o valor da variável dependente.

Vetores de suporte: Amostras críticas em SVM/SVR que definem o limite de decisão ou a função de regressão.

RESUMO

A cultura da soja desempenha papel fundamental no agronegócio brasileiro, exigindo práticas de manejo cada vez mais eficientes para garantir altos níveis de produtividade e sustentabilidade. Entretanto, os critérios atualmente utilizados pelos agrônomos para recomendar doses de fertilizantes e corretivos apresentam elevada divergência, consequência direta da escassez de pesquisas de calibração regionalizadas que considerem, simultaneamente, atributos específicos do solo, condições climáticas, altitude, sistema de plantio e demais variáveis que influenciam o desenvolvimento da planta. Essa limitação dificulta a tomada de decisão técnica no campo e reduz a precisão das recomendações agronômicas.

Nesse contexto, torna-se essencial compreender a relação entre esses fatores e a produtividade, de modo a permitir estimativas mais robustas e apoiar decisões de manejo. O avanço das técnicas de Inteligência Artificial (IA) oferece novas possibilidades para modelar tais relações e identificar padrões complexos, muitas vezes imperceptíveis aos métodos tradicionais. Diante disso, este trabalho apresenta o desenvolvimento do software Guiafertil, uma ferramenta computacional voltada à análise exploratória de dados agronômicos e à predição do limiar de produtividade da soja. O Guiafertil integra banco de dados, rotinas de pré-processamento, análise estatística, geração de gráficos e aplicação de algoritmos de aprendizado de máquina — incluindo Regressão Linear, Máquinas de Vetores de Suporte (SVR), Floresta Aleatória e Redes Neurais Artificiais. Utilizando dados reais provenientes dos "Cases Campeões" do Comitê Estratégico Soja Brasil (CESB), o sistema permite avaliar atributos de solo, clima e manejo, treinar modelos, analisar métricas de desempenho e estimar a produtividade a partir de cenários práticos, como diferentes doses de adubação. Como principais resultados, destaca-se que a Rede Neural (MLP - Multi-Layer Perceptron) apresentou o melhor desempenho preditivo entre os modelos avaliados, alcançando desempenho R^2 igual a 0,56 em termos de predição sacas/ha, enquanto a Regressão Linear mostrou limitações significativas em termos preditivos com R^2 igual a -0,60. Além disso, neste estudo de caso demonstrou-se que ajustes nas doses de fertilizantes alteram os custos de produção, mas têm impacto limitado na produtividade prevista, evidenciando a importância de decisões mais assertivas. A qualidade e usabilidade do software foram avaliadas por engenheiros agrônomos, que atribuíram notas médias superiores a 9,0 para a maior parte dos critérios analisados. Assim, o Guiafertil se mostra uma ferramenta promissora para apoiar a recomendação agronômica, contribuindo para o uso mais eficiente de insumos, redução de custos e maior sustentabilidade no cultivo da soja. O trabalho também abre caminho para pesquisas futuras envolvendo novas culturas, ampliação da base de dados e integração com tecnologias de agricultura de precisão.

Palavras-Chave: Cultura da Soja; Floresta Aleatória; Produtividade; Regressão Linear; Rede Neural; SVR

ABSTRACT

Soybean cultivation plays a fundamental role in Brazilian agribusiness, requiring increasingly efficient management practices to ensure high levels of productivity and sustainability. However, the criteria currently used by agronomists to recommend fertilizer and soil amendment rates exhibit considerable variability, mainly due to the lack of regional calibration studies that simultaneously consider specific soil attributes, climatic conditions, altitude, planting systems, and other variables that influence crop development. This limitation hinders technical decision-making in the field and reduces the accuracy of agronomic recommendations.

In this context, understanding the relationship between these factors and crop productivity becomes essential to enable more robust estimates and support management decisions. Advances in Artificial Intelligence (AI) techniques offer new opportunities to model such relationships and identify complex patterns that are often difficult to detect using traditional approaches. Therefore, this study presents the development of Guiafertil, a computational tool designed for exploratory analysis of agronomic data and soybean yield threshold prediction. Guiafertil integrates a database management system, data preprocessing routines, statistical analysis, graphical visualization, and machine learning algorithms, including Linear Regression, Support Vector Regression (SVR), Random Forest, and Artificial Neural Networks.

Using real-world data obtained from the "Champion Cases" database of the Brazilian Soybean Strategic Committee (CESB), the system enables the evaluation of soil, climate, and management attributes, model training, performance analysis, and productivity estimation based on practical scenarios, such as different fertilization rates. Among the main results, the Multi-Layer Perceptron (MLP) Neural Network achieved the best predictive performance, reaching an R^2 value of 0.56 for soybean yield prediction bags/ha, whereas Linear Regression showed significant predictive limitations, obtaining an R^2 value of -0.60. Furthermore, the case study demonstrated that adjustments in fertilizer application rates affect production costs but have limited impact on predicted productivity, highlighting the importance of more accurate decision-making processes.

The software quality and usability were evaluated by agronomists, who assigned average scores above 9.0 for most of the assessed criteria. Therefore, Guiafertil proved to be a promising tool for supporting agronomic recommendations, contributing to more efficient use of agricultural inputs, cost reduction, and greater sustainability in soybean production. Additionally, this work opens new opportunities for future research involving other crops, expansion of the database, and integration with precision agriculture technologies.

Keywords: Linear Regression; Neural Network; Productivity; Soybean Cultivation; Random Forest; SVR

1 INTRODUÇÃO

Solos com características químicas inadequadas, tais como: elevada acidez, altos teores de alumínio trocável e deficiência de cálcio e magnésio, entre outras, afetam a eficiência do manejo agrícola (RIBEIRO, 1999). Para as plantas, os maiores prejuízos causados por solos ácidos resumem-se na fitotoxicidade do alumínio e do manganês e pela deficiência de cálcio e magnésio no solo (REATTO, 2004). Por outro lado, uma vez identificados e corrigidos esses problemas, o solo poderá apresentar grande potencial agrícola (RIBEIRO, 1999).

Ribeiro (1999) também destaca a necessidade de pesquisas de "calibração" que indiquem, com maior precisão, a quantidade ideal de insumos agrícolas a ser aplicada, considerando atributos específicos.

A cultura da soja desempenha papel estratégico no agronegócio brasileiro, contribuindo significativamente para a economia nacional e para o abastecimento global. A crescente demanda por produtividade, aliada à busca por práticas agrícolas mais sustentáveis, torna cada vez mais necessário o uso eficiente de insumos, correção do solo e recomendações de manejo baseadas em evidências técnicas. No entanto, observa-se elevada divergência entre as recomendações fornecidas por profissionais da área, sobretudo devido à escassez de pesquisas de calibração regionalizadas que integrem de forma simultânea atributos do solo, clima, altitude, sistema de plantio, manejo nutricional e demais fatores que influenciam o rendimento da cultura.

Para orientar o manejo adequado da soja, agrônomos e consultores precisam responder a questões essenciais relacionadas às condições ideais para alcançar a produtividade esperada, tais como: quais são os níveis adequados de nutrientes (N, P, K, Ca, Mg, S, Zn, B, Cu, Fe, Mn, Mo, Ni, Co), matéria orgânica (MO), carbono orgânico (CO), pH, saturação por alumínio e saturação de bases; qual é a faixa ideal de altitude, temperatura mínima e máxima, e precipitação pluviométrica; e quais são as doses ideais de fertilizantes e corretivos a serem aplicados. Essas perguntas são fundamentais para estimar a produtividade, definir a viabilidade econômica do cultivo e orientar estratégias de manejo. Contudo, a ausência de métodos consolidados que integrem tais variáveis de forma sistemática dificulta o processo de tomada de decisão e pode comprometer o desempenho das lavouras.

O desenvolvimento de ferramentas tecnológicas que permitam interpretar grandes volumes de dados agronômicos de forma integrada surge como uma alternativa promissora para superar essas limitações. A Inteligência Artificial (IA), especialmente por meio de técnicas de

aprendizado de máquina, oferece potencial para identificar padrões complexos e relações não lineares entre variáveis que influenciam a produtividade da soja. A literatura demonstra que modelos de aprendizado supervisionado podem reconhecer padrões ocultos nos dados e utilizá-los para prever resultados futuros com maior precisão, ampliando a capacidade analítica e decisória dos profissionais envolvidos no processo produtivo.

Neste contexto, este trabalho apresenta o desenvolvimento do software Guiafertil, uma ferramenta computacional destinada à análise exploratória de dados agronômicos e à predição do limiar de produtividade da soja. O sistema integra banco de dados, rotinas de pré-processamento, geração de gráficos, análise estatística e aplicação de algoritmos de aprendizado de máquina, incluindo redes neurais artificiais, Regressão Linear, Máquinas de Vetores de Suporte (SVM) e Floresta Aleatória. A utilização de dados reais provenientes dos "Cases Campeões" do CESB torna possível a construção de modelos baseados em situações práticas, aproximando o software da realidade do campo.

A hipótese central deste estudo é que modelos de aprendizado de máquina são capazes de identificar padrões complexos entre atributos do solo, clima, altitude, sistema de plantio e manejo nutricional, permitindo prever a produtividade da soja com maior precisão do que métodos tradicionais. Acredita-se que a aplicação dessas técnicas pode oferecer suporte técnico mais assertivo, contribuindo para decisões de manejo mais eficientes e sustentáveis.

O objetivo geral deste trabalho é desenvolver um software baseado em IA capaz de analisar atributos agronômicos e estimar a produtividade da cultura da soja, fornecendo apoio técnico ao processo de recomendação de insumos agrícolas. Para isso, foram estabelecidos objetivos específicos, tais como: desenvolver modelos de aprendizado de máquina; implementar rotinas de pré-processamento e análise exploratória; criar uma interface gráfica intuitiva composta por módulos de cadastro, gerenciamento de dados agronômicos e ambiente de predição; integrar ferramentas estatísticas e métricas de desempenho; realizar estudo de caso para análise do impacto de variações na adubação sobre a produtividade prevista; e validar o software por meio da avaliação de engenheiros agrônomos.

A justificativa para o desenvolvimento deste trabalho fundamenta-se na necessidade crescente de ferramentas tecnológicas que auxiliem o setor agrícola na utilização eficiente de insumos, reduzindo custos, aumentando a precisão das recomendações e promovendo a sustentabilidade produtiva. A ausência de sistemas integrados que utilizem IA para análise e

predição da produtividade representa uma lacuna relevante na prática agronômica. Ao oferecer uma solução prática e acessível, o software Guiafertil contribui para modernizar a agricultura, proporcionando ao produtor e ao consultor informações confiáveis e embasadas em dados reais.

Neste capítulo foi apresentado o contexto da pesquisa, o problema enfrentado na prática agronômica, a justificativa, os objetivos e a hipótese que fundamentam o estudo. Os capítulos seguintes estão organizados da seguinte forma: o Capítulo 2 apresenta o referencial teórico, abordando os fundamentos da cultura da soja e os principais conceitos relacionados às técnicas de aprendizado de máquina utilizadas no estudo, incluindo Redes Neurais, Regressão Linear, Máquinas de Vetores de Suporte e Floresta Aleatória. O Capítulo 3 descreve a metodologia adotada, detalhando a base de dados, as etapas de pré-processamento, a configuração dos modelos e a arquitetura do software Guiafertil. O Capítulo 4 apresenta os resultados obtidos, incluindo a avaliação do desempenho dos modelos preditivos, análises comparativas e estudo de caso com diferentes cenários de adubação. Por fim, o Capítulo 5 apresenta as conclusões do trabalho, destacando as principais contribuições, limitações e sugestões para pesquisas futuras.

.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta o referencial teórico que fundamenta o desenvolvimento deste trabalho, abordando tanto os aspectos agronômicos relacionados à cultura da soja quanto os conceitos e técnicas de aprendizado de máquina aplicados à predição de produtividade. Inicialmente, são discutidos os principais fatores que influenciam o desempenho produtivo da cultura, incluindo atributos do solo, condições climáticas e práticas de manejo. Em seguida, são apresentados os fundamentos dos algoritmos utilizados neste estudo, como Redes Neurais Artificiais, Regressão Linear, Máquinas de Vetores de Suporte e Floresta Aleatória, destacando seus princípios, aplicações e limitações. Por fim, são discutidos estudos relacionados que utilizam técnicas semelhantes.

2.1 CULTURA DA SOJA

O Brasil, pela sua vasta extensão territorial, apresenta uma ampla diversidade de solos, climas e sistemas de produção, o que faz com que a cultura da soja manifeste comportamentos agronômicos distintos entre as diferentes regiões produtoras do país. Embora os dados analisados neste estudo incluam lavouras de todo o território nacional, abordar simultaneamente todas essas particularidades tornaria o referencial teórico excessivamente extenso e disperso, fugindo ao propósito desta seção. Assim, para fins de fundamentação agronômica, optou-se por adotar a região do Cerrado como referência principal, por se tratar de uma das regiões mais representativas da sojicultura brasileira. Compreender a soja sob as condições típicas do Cerrado permite discutir, de forma clara e objetiva, as interações entre solo, clima, nutrição mineral e manejo, estabelecendo uma base sólida para sustentar as análises desenvolvidas neste trabalho, sem restringir o alcance geográfico dos dados utilizados.

A compreensão da produtividade da soja depende do entendimento integrado dos fatores agronômicos que influenciam o desenvolvimento da planta, sobretudo em regiões do Cerrado, onde predominam solos altamente intemperizados e de baixa fertilidade natural. Esses fatores incluem as características químicas e físicas do solo, a disponibilidade de nutrientes essenciais, as condições climáticas, o sistema de plantio e o manejo da fertilidade do solo. O estudo desses elementos é fundamental para embasar recomendações de manejo e interpretar corretamente a resposta da cultura em diferentes ambientes, conforme destacado por Reatto et al. (2024).

A soja apresenta elevado potencial produtivo, mas sua expressão depende de condições favoráveis ao longo do ciclo. Entre os fatores mais determinantes estão temperatura, disponibilidade hídrica, radiação solar e fertilidade do solo. No Cerrado, a distribuição sazonal de chuvas, marcada por períodos de déficit hídrico, e a variação térmica durante o ciclo são elementos que afetam diretamente o desempenho da cultura. Reatto et al. (2024) explicam que a água disponível no solo e sua capacidade de retenção estão diretamente relacionadas à textura, ao teor de matéria orgânica e à estrutura do solo, influenciando a absorção de nutrientes e o crescimento radicular.

Em relação ao solo, os Latossolos constituem a classe predominante no Cerrado. Eles são profundos, bem drenados e apresentam elevada estabilidade estrutural, características que favorecem o desenvolvimento do sistema radicular da soja. No entanto, possuem baixa saturação por bases, altos teores de alumínio trocável e reduzida disponibilidade natural de fósforo — limitações que exigem correção adequada para garantir produtividade. Segundo Reatto et al. (2024), a acidez elevada compromete o crescimento das raízes e reduz a eficiência na absorção de nutrientes, especialmente fósforo, potássio, cálcio e magnésio. A prática da calagem, portanto, é essencial para elevar o pH, reduzir o alumínio tóxico e aumentar a disponibilidade de cálcio e magnésio no perfil do solo.

A gessagem complementa a calagem ao fornecer cálcio e enxofre em profundidade, favorecendo o desenvolvimento radicular além da camada superficial. Reatto et al. (2024) destacam que, em solos de Cerrado, o gesso agrícola melhora a estrutura química do subsolo, reduz a saturação por alumínio, e o que aumenta a tolerância das plantas a períodos de estiagem, uma vez que uma maior profundidade radicular maior está associada à maior absorção de água.

Os nutrientes desempenham papel central no crescimento e na produtividade da cultura. Entre os macronutrientes, destacam-se nitrogênio, fósforo, potássio, cálcio, magnésio e enxofre. O nitrogênio, embora essencial, é suprido principalmente por meio da fixação biológica, que depende de condições adequadas de pH e do teor de cálcio. O fósforo, geralmente limitado em solos do Cerrado devido à fixação por óxidos de ferro e alumínio, deve ser aplicado em doses suficientes para suprir a demanda da cultura. O potássio, fortemente influenciado pela textura do solo, apresenta maior mobilidade em solos arenosos, exigindo manejo apropriado para evitar perdas por lixiviação. Reatto et al. (2024) enfatizam que a interpretação da análise de solo é imprescindível para o ajuste adequado da adubação.

Os micronutrientes, apesar de necessários em pequenas quantidades, são essenciais ao desenvolvimento da soja. Elementos como zinco, manganês, cobre, ferro, boro e molibdênio influenciam processos metabólicos fundamentais, e sua deficiência pode comprometer severamente a produtividade. As condições típicas dos solos do Cerrado como baixa matéria orgânica e pH muito elevado após calagem podem reduzir a disponibilidade desses micronutrientes. Isso reforça a importância do monitoramento por meio de análises químicas e também o uso de análises foliares, visando a aplicações suplementares de adubos foliares.

As condições climáticas também afetam diretamente a fisiologia da soja. A cultura possui faixas de temperatura ideais para germinação, desenvolvimento vegetativo e fase reprodutiva. Temperaturas extremas, tanto baixas quanto elevadas, podem comprometer a formação de flores, o enchimento de grãos e a eficiência do uso da água. A precipitação durante o ciclo é outro fator crucial, sendo a fase reprodutiva a mais sensível ao déficit hídrico. Reatto et al. (2024) observam que a interação entre textura, matéria orgânica e estrutura do solo determina sua capacidade de armazenamento de água, influenciando diretamente a resiliência da cultura em períodos de menor precipitação.

Por fim, o sistema de plantio influencia tanto a fertilidade quanto as condições físicas do solo. O plantio direto, amplamente utilizado no Cerrado, contribui para a conservação da umidade, o aumento da matéria orgânica e a redução da erosão. Essas características tornam este sistema uma estratégia eficaz para mitigar limitações edafoclimáticas e aumentar a estabilidade produtiva. A adoção de práticas conservacionistas, aliada ao manejo adequado da fertilidade do solo, é essencial para que a soja expresse seu potencial produtivo em ambientes de Cerrado.

Desse modo, a literatura agronômica evidencia que a produtividade da soja depende de um conjunto de fatores que atuam de forma integrada. O entendimento desses fundamentos sobre solo, clima, nutrição mineral e manejo agrícola conforme apresentados por Reatto et al. (2024) é indispensável para interpretar os resultados obtidos neste estudo e fundamentar as análises subsequentes relacionadas à cultura da soja.

Técnicas de mineração de dados e aprendizado de máquina, como Rede Neural Artificial, Regressão Linear, Máquina de Vetores de Suporte (SVM) e Floresta Aleatória, permitem modelar relações complexas entre variáveis agronômicas e a produtividade da soja. Esses métodos identificam padrões e interações relevantes, possibilitando a construção de modelos preditivos que auxiliam a tomada de decisão no contexto agrícola.

2.2 DESCOBERTA DE CONHECIMENTO EM DADOS

Não adianta o produtor de soja ter uma grande quantidade de dados sobre produtividade, análise de solo, recomendações de adubação etc., tem pouca utilidade para o produtor rural se ele não estiver satisfeito com os resultados obtidos na sua lavoura. A decisão para alcançar uma produtividade economicamente vantajosa deve considerar uma abordagem sofisticada e eficiente. Nesse sentido, o uso de mineração de dados e técnicas de IA pode ser uma alternativa eficaz para gerar informações e conhecimentos valiosos.

Uma tabela em banco de dados, é composta por linhas e colunas, já em aprendizado de máquina, cada coluna é um atributo, cada linha é uma instancia (registro) e cada tabela uma relação. Por exemplo, o teor de matéria orgânica (atributo) é uma das características da análise de solo (relação) e seu valor (instancia) que pode ser comparada com outro fato no sistema (AMARAL, 2016).

A classificação, que é uma tarefa muito comum no aprendizado de máquinas, tem como objetivo usar todos os atributos que compõe uma relação para prever um atributo especial (AMARAL, 2016). Por exemplo por meio dos atributos do solo, recomendações de insumos e produção agrícola coletados ao longo do tempo é possível prever o índice de produtividade (atributo especial).

Em mineração de dados existem grandes grupos de dados, como mostra a Tabela 1 abaixo:

TABELA 1 . Classificação dos tipos de dados e suas características. Fonte: Elaborado pelo autor.

Tipo de Dado	**Qualitativa** (atributo categórico) Classifica por tipo / atributo.	**Nominal:** Características / atributos que **não podem ser ordenados.** Exemplo: Sistema de plantio: Plantio direto e plantio convencional.
		Ordinal: Características / atributos que **podem ser ordenados.** Exemplo: Disponibilidade de fosforo (P) no solo: Baixa, Média e Alta.
	Quantitativa (atributo numérico) Escala de classificação numérica	**Continua: Entre dois pontos** de escala existe **infinitos valores.** Exemplo: Análise química do solo como: K +, Ca 2+, Mg 2+ (mg dm-3).
		Discreta: Entre dois pontos de escala existe **número finito de valores.** Exemplo: Numero de analises de solo avaliadas.

As tarefas de mineração de dados podem ser organizadas em quatro grupos (AMARAL, 2016):

1. Classificação: Utiliza todos os atributos que compõe uma relação para prever um atributo especial (classe).
2. Regressão: É um tipo de classificação, utiliza atributos de entrada já observados para prever uma resposta, neste caso, procura-se estimar um valor numérico e não uma classificação.
3. Agrupamento: Não existe classe, é criado grupos a partir das características, ou atributos instanciados semelhantes entre os elementos. Dependendo do tipo de agrupamento um elemento, pode pertencer a mais de um grupo ou não ser agrupado, ou seja, ser considerado um ruido (erro).
4. Regras de associação: Encontrar relacionamentos ou padrões frequentes entre conjuntos de dados. Exemplo: A capacidade que das plantas de absorver os nutrientes do solo está intimamente ligado ao seu pH. (ideal entre 5,5 a 6,3) (REATTO, 2004).

A classificação é uma tarefa supervisionada, ou seja, existe uma classe, ou um atributo especial com o qual se pode comparar e validar resultado. Já em uma tarefa não supervisionado não existe a classe, não há um rotulo prévio. Agrupamentos são exemplos de tarefas não supervisionadas.

Cada tipo de tarefa tem um objetivo especifico que pode ser implementado por diferentes tipos de algoritmos como mostra a Tabela 2 abaixo:

TABELA 2 . Classificação das tarefas de mineração de dados, tipos de algoritmos e exemplos. Fonte: Elaborado pelo autor.

Tarefas	**Tipos de Algoritmos**	**Algoritmos**
Classificação	Bayes	Naive Bayes
		BayesNet
	Rules	OneR
	Redes Neurais	Backpropagation
Agrupamentos	Por Densidade	DBSCAN
	Centroide	K-means
Regras de Associação	Frequência de Subconjuntos	Apriori
	Árvore de dados	FP Growth

Este estudo emprega técnicas de aprendizado de máquinas como Rede Neural, Regressão Linear, Floresta Aleatória e SVM para compreender as interações entre atributos do solo, clima, altitude, sistema de plantio, e fertilizantes com o objetivo de estimar a produtividade na cultura da soja.

2.3 REDE NEURAL

Grus (2021) destaca a Rede Neural Artificial como um modelo preditivo baseado na dinâmica do cérebro, que tem uma série de neurônios conectados. Cada um deles analisa as saídas dos outros neurônios ligados nele, faz um cálculo e, em seguida, irá disparar (se o valor calculado exceder um limite) ou não disparar (se não exceder um limite).

Géron (2021) define a composição de uma Rede Neural Profunda como: “Uma MLPs (Perceptrons Multicamadas) é composta de uma ou mais camadas de TLUs (Unidades Lógicas de Limite), chamadas de camadas ocultas, e uma camada final de TLUs, chamada de camada de saída”.

A Figura 1 mostra um exemplo de MLP.

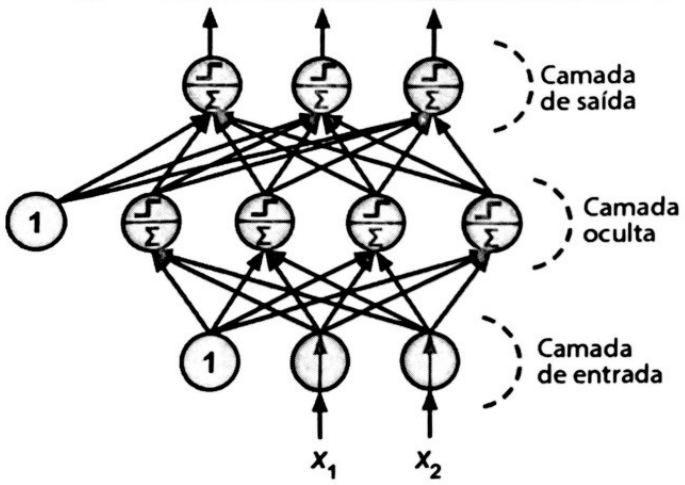


FIGURA 1 . Arquitetura de um perceptron multicamada, destacando as camadas de entrada, oculta e saída, bem como as conexões totalmente interligadas entre os neurônios.
Fonte: GÉRON (2021).

Antes de serem processados pela Rede Neural, os dados passam por um pré-processamento, com o objetivo de melhorar sua qualidade. Após o pré-processamento, os dados

são enviados para os neurônios da camada de entrada, que estão conectados a outros neurônios na camada oculta por meio de seus respectivos pesos (sinapses). Nessa camada, ocorre o processamento dos dados, que podem atravessar várias camadas interligadas, também com seus pesos ajustados. Após o processamento, uma ou mais respostas são geradas e encaminhadas para a camada de saída. Os neurônios são ativados por funções de ativação, que produzem uma resposta única. Neste estudo a função ReLU é utilizada nas camadas ocultas e a função linear na camada de saída.

Função de ativação ReLU (*Rectified Linear Unit*). Transforma a saída de cada neurônio como:

$$f(x) = max\ (0, x) \quad (1)$$

Onde:

Para entradas positivas: *(x > 0)*: A saída é igual à entrada *f(x) = x*.

Para entradas negativas: *(x ≤ 0)*: A saída é truncada para zero *f(x) = 0*.

x: é o resultado da operação linear no neurônio, que combina os pesos, as entradas e o viés do neurônio.

$$x = \sum_{i=1}^{n} (w_i . x_i) + b \quad (2)$$

Onde:

wi : Peso associadoa i-ésima entrada do neurônio.

xi : Valor da i-ésima entrada (proveniente da camada anterior ou diretamente do conjunto de dados).

b : Viés (bias) do neurónio, que permite ajustes independentes dos pesos.

n : Número de entradas no neurônio.

Na função de ativação linear na camada de saída é diretamente proporcional à entrada, expressa como:

$$f(x) = x \quad (3)$$

Onde:

A saída da função *f(x)* é igual à entrada, sem transformações.

Inicialmente os pesos da Rede Neural são gerados aleatoriamente, e posteriormente evoluídos (encontrar os pesos ideais) durante o processo de aprendizado.

Géron (2021) destaca que as MLPs podem ser utilizadas para tarefas de regressão. Para prever um único valor, é necessário ter um neurônio de saída para cada dimensão da saída. A função de perda usada no treinamento é geralmente o erro médio quadrático. Entretanto, se o conjunto de treinamento contiver muitos outliers, o erro absoluto médio pode ser uma escolha melhor.

2.4 REGRESSÃO LINEAR

Mueller (2020) define a Regressão Linear como uma reta que passa por uma série de coordenadas x/y, determinando a localização de um ponto de dados. Os pontos de dados nem sempre estão diretamente na reta, mas ela mostra onde estariam, em um mundo perfeito de coordenadas lineares. Com a reta, é possível predizer o valor de y, tendo o valor de x determinado. Quando há apenas uma variável preditora, temos uma Regressão Linear Simples e quando há muitas, temos uma Regressão Linear Múltipla.

Um modelo de Regressão Linear realiza predições calculando uma soma ponderada das características de entrada, somada a uma constante conhecida como viés (também chamada de intercepto ou coeficiente linear), conforme ilustrado na equação a seguir (GÉRON, 2021):

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \quad (4)$$

Onde:

$\hat{y}$ é o valor previsto.

n é o número das características.

x_i é o valor da i-ésima característica.

θ_j é o j-ésimo parâmetro do modelo (incluindo o viés θ_0 e os pesos das características θ_1, θ_2, ... θ_n).

O treinamento de um modelo de Regressão Linear requer a determinação do valor de θ que minimiza a raiz do erro quadrático médio (RMSE). Conforme apontado por Géron (2021), a equação normal é uma solução eficaz para calcular o valor de θ que reduz o erro ao mínimo. A fórmula matemática dessa equação é apresentada a seguir:

$$\hat{\theta} = (X^T X)^{-1} X^T y \quad (5)$$

Onde:

$\hat{\theta}$: é o valor de θ que minimiza a erro.

X : é a matriz de características, contendo os valores de entrada. Cada linha representa uma observação e cada coluna, uma característica.

X^T : é a transposta da matriz X.

$(X^T X)^{-1}$: é a inversa do produto de X^T e X. Esse termo é crucial para determinar os coeficientes e exige que seja $X^T X$ invertível.

y é o vetor de valores-alvo contendo $y^{(1)}$ a $y^{(m)}$.

2.5 MÁQUINA DE VETORES DE SUPORTE (SVM)

SVM, Support Vector Machine (Máquina de Vetores de Suporte), é uma técnica de aprendizado de máquina amplamente aplicada em tarefas de classificação e regressão. Este modelo supervisionado visa identificar um hiperplano ótimo que separe os dados em classes distintas ou forneça aproximações precisas para valores contínuos, no caso de regressão. As Figuras 2 e 3 apresentadas por Géron (2021) ilustra claramente o conceito central da regressão utilizando SVM como modelo preditivo.

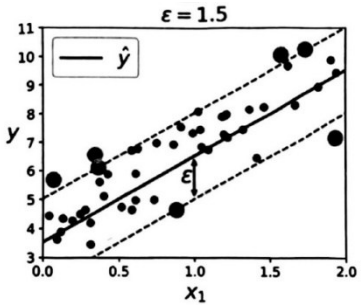


FIGURA 2 . Exemplo de regressão SVM linear, destacando a função de predição e a margem ε-insensitive. Fonte: GÉRON (2021).

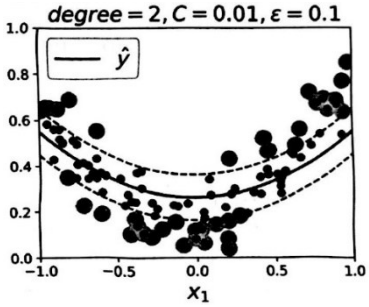


FIGURA 3 . Exemplo de regressão SVM não linear (kernel polinomial), ilustrando a curva ajustada e sua margem ε-insensitive. Fonte: GÉRON (2021).

A regressão por SVM, é também conhecida como SVR Support Vector Regression (Regressão de Vetores de Suporte), utilizada para modelar relações entre variáveis em cenários de predição. Diferentemente da Regressão Linear tradicional, a SVR busca não apenas ajustar uma linha que represente os dados, mas também introduz o conceito de margem de tolerância (ε), que define um intervalo ao redor da linha de regressão dentro do qual os desvios não são penalizados. Na figura 2 apresentado, a linha sólida ($\hat{y}$) representa a função predita pela SVR, que tenta capturar a relação entre a variável dependente y e a variável independente X_1. Ao redor dessa linha, encontram-se duas linhas pontilhadas paralelas que delimitam a margem de tolerância, geradas pelo épsilon (ε). Essa margem define o intervalo em que os desvios entre os valores observados e os valores preditos são considerados aceitáveis, sem impacto na função objetivo do modelo. Os pontos no gráfico representam as observações. Aqueles que se encontram dentro da margem de tolerância são tratados como suficientemente próximos à predição e, portanto, não influenciam na otimização do modelo. Em contraste, os pontos que estão fora dessa margem de tolerância são chamados de vetores de suporte, pois desempenham um papel fundamental na definição da linha de regressão. Já a figura 3 ilustra a aplicação da SVR a um problema com uma relação não-linear entre a variável dependente (y) e a variável independente (X_1) utilizado um kernel polinomial de grau 2, conforme indicado pela anotação "degree = 2", para modelar a relação curvilínea observada nos dados. A SVR é uma técnica poderosa que permite capturar padrões complexos ao mapear os dados para um espaço dimensional mais alto, tornando possível a construção de modelos não-lineares. A linha sólida no gráfico ($\hat{y}$) representa a função ajustada pelo modelo SVR, enquanto as linhas pontilhadas acima e abaixo da curva delimitam a margem de tolerância (ε=0.1). O parâmetro $C=0.01$ desempenha um papel no equilíbrio entre a suavidade do modelo e sua capacidade de ajuste aos dados. Um valor baixo de C, como neste caso, indica que o modelo prioriza uma solução mais simples e robusta, permitindo que alguns desvios maiores ocorram sem afetar drasticamente a função de otimização. Isso torna o modelo menos suscetível a outliers, favorecendo sua generalização para novos dados. Aqueles pontos dentro da margem de tolerância (ε) não influenciam diretamente o ajuste do modelo, enquanto os pontos fora da margem impactam a construção da curva, forçando o modelo a levar em conta os desvios mais significativos.

2.6 FLORESTA ALEATÓRIA

Géron (2021) explica que as respostas agregadas geralmente superam a resposta de um único especialista, fenômeno conhecido como *sabedoria das multidões*. Nesse contexto, um grupo de preditores tende a produzir resultados melhores do que o melhor preditor individual. Esse conjunto de predições é conhecida como método ensemble. A Floresta aleatória é um algoritmo de aprendizado de máquina baseado em um conjunto de árvores de decisão. Ele combina múltiplas árvores para melhorar a precisão e a generalização do modelo. Esse método faz parte da categoria dos métodos ensemble e pode ser usado tanto para classificação quanto para regressão.

O funcionamento do modelo pode ser dividido em três etapas principais: criação de subconjuntos de dados, construção das árvores de decisão e combinação dos resultados.

- Criação de Subconjuntos de Dados: é gerado diferentes subconjuntos do conjunto de dados de treinamento. Cada árvore é treinada de forma independente utilizando um desses subconjuntos, o que garante que as árvores sejam variadas e menos propensas a erros sistemáticos.
- Construção das Árvores de Decisão: Durante a construção das árvores de decisão, um aspecto importante da Floresta Aleatória é a seleção aleatória de variáveis em cada divisão dos nós. Em vez de considerar todas as variáveis disponíveis, apenas um subconjunto aleatório é analisado. Isso introduz ainda mais aleatoriedade, reduzindo o risco de correlação entre as árvores e aumentando a capacidade de generalização do modelo.
- Combinação dos Resultados: A previsão final da Floresta Aleatória é baseada na combinação das predições individuais de todas as árvores. Na classificação o resultado é determinado pelo voto da maioria das árvores e na regressão a previsão final é a média das predições de todas as árvores.

A Figura 4 ilustra uma Floresta Aleatória na predição de produtividade em sacas para a cultura da soja.

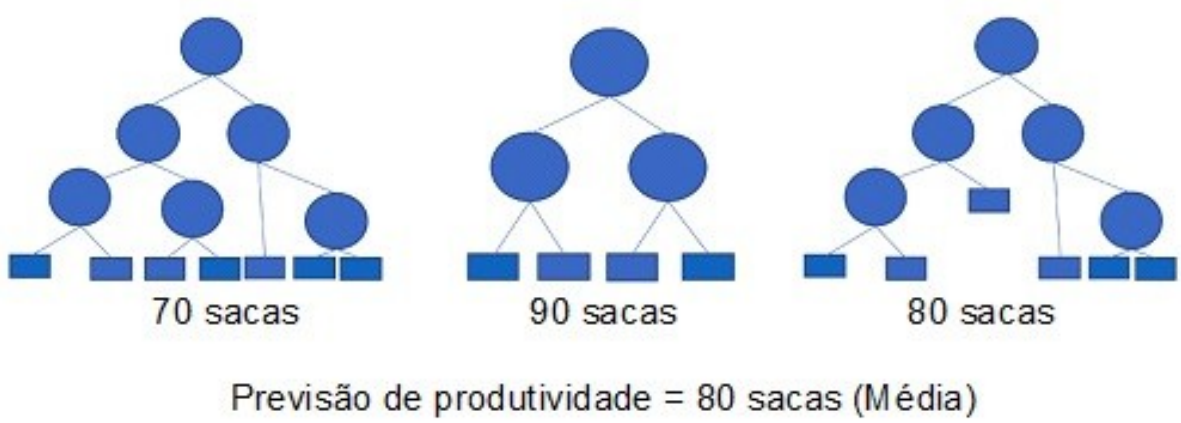


FIGURA 4 . Funcionamento da Floresta Aleatória na estimativa de produtividade, mostrando a combinação das previsões das árvores. Fonte: Elaborado pelo autor.

2.7 ESTUDOS RELACIONADOS

O uso de técnicas de Aprendizado de Máquina (Machine Learning) na agricultura tem se expandido rapidamente, especialmente para tarefas de predição de produtividade. Os estudos disponíveis na literatura demonstram avanços importantes, mas também revelam lacunas que justificam a proposta do software Guiafertil, que busca integrar múltiplos fatores agronômicos em um único ambiente preditivo.

Shammi (2024) empregou imagens multiespectrais obtidas por VANT (Veículo Aéreo Não Tripulado.) para monitorar o crescimento da soja e prever sua produtividade utilizando diferentes algoritmos de Aprendizado de Máquina. O estudo demonstra que variáveis espectrais capturam bem o vigor vegetativo, mas dependem de infraestrutura tecnológica específica, restringindo sua aplicabilidade para produtores sem acesso a drones ou sensores avançados

Santos (2024) analisou o desempenho da Floresta Aleatória na previsão da produtividade da soja sob diferentes sistemas de culturas de cobertura. Os dados foram obtidos em experimentos agronômicos tradicionais, concentrando-se nas relações entre manejo de cobertura e rendimento. Embora o modelo tenha apresentado desempenho satisfatório, o estudo é limitado ao contexto experimental e não integra múltiplas variáveis de solo e clima simultaneamente, como proposto neste trabalho

Joshi (2023) utilizou sensoriamento remoto por satélite para desenvolver modelos de IA voltados à previsão de rendimento da soja. O enfoque recai sobre índices espectrais derivados de

imagens orbitais, reforçando o papel dessas variáveis na detecção de estresse e na estimativa de produtividade. Entretanto, o estudo não considera atributos edáficos diretamente, o que reduz a capacidade de modelar efeitos relacionados à fertilidade do solo

Pejak (2022) combinou dados multiespectrais do Sentinel-2 com características do solo para treinar modelos de aprendizado de máquina em escala intra-talhão. O estudo se destaca por integrar informações edáficas e espectrais, demonstrando que a inclusão de atributos do solo melhora significativamente a precisão dos modelos. Todavia, o trabalho tem foco restrito à escala espacial fina, não abordando fatores climáticos ou de manejo como doses de fertilizantes

Khalila (2021) avaliou modelos de Redes Neurais aplicados à estimativa de produtividade de grãos a partir de variáveis multiespectrais. O estudo comparou arquiteturas e fontes de dados, evidenciando que redes neurais capturam não linearidades relevantes na variabilidade espacial das culturas. No entanto, assim como outros trabalhos baseados somente em sensoriamento remoto, a ausência de atributos agronômicos mais amplos limita a capacidade de interpretar causas fisiológicas da produtividade

Guimarães (2019) realizou uma comparação entre modelos de aprendizado de máquina — incluindo MLP, Floresta Aleatória e XGBoost e um modelo agrometeorológico convencional para predição da produtividade de soja. Os resultados indicaram melhor desempenho dos algoritmos de IA, reforçando a viabilidade de abordagens não lineares para esse tipo de tarefa. Entretanto, o estudo utiliza apenas dados de clima e solo, sem considerar informações de manejo, como doses de fertilizantes, que são essenciais para recomendações práticas no campo

Leal et al. (2015) investigaram a eficácia de Redes Neurais na predição de produtividade do milho segunda safra, utilizando atributos do solo em ambiente de Cerrado. Os resultados mostraram que as RNAs superam modelos de regressão na captura de relações complexas entre textura, CTC, matéria orgânica e rendimentos agrícolas. O estudo reforça a importância da variabilidade edáfica como fator-chave da produtividade, mas não incorpora variáveis climáticas ou de manejo

Soares et al. (2015) também aplicaram Redes Neurais para estimar a produtividade do milho com base em características morfológicas das plantas. Embora os modelos tenham apresentado alta capacidade preditiva, o enfoque permanece restrito a um único grupo de variáveis, não oferecendo um panorama integrador como o do presente trabalho

Vriesmann et al. (2004) compararam Árvore de Decisão e Algoritmos Genéticos na mineração de dados sobre produtividade da soja e atributos físico-químicos do solo. O estudo demonstrou que abordagens evolutivas podem identificar padrões relevantes em bases de dados complexas, com desempenho superior no tratamento de variáveis contínuas. Porém, trata-se de uma análise estática, focada na geração de regras, e não em modelos de predição aplicáveis a diferentes cenários agronômicos

Por fim, Bucene e Rodrigues (2004) utilizaram Redes Neurais para classificar a "fertilidade aparente" do solo com base em atributos como pH, CTC, saturação por bases e teores de macronutrientes. Embora não trate diretamente de produtividade, o estudo é relevante por demonstrar a eficiência das RNAs em aplicações agronômicas que envolvem variáveis químicas do solo. Ainda assim, a abordagem é limitada à classificação e não envolve predição contínua de rendimento ou integração com dados climáticos e de manejo.

Comparando esses trabalhos com o software Guiafertil, observa-se que cada estudo aborda apenas uma parte do problema sejam dados espectrais, atributos do solo, clima, manejo ou morfologia da planta. O diferencial deste projeto está em:

- Integrar simultaneamente solo, clima, altitude, sistema de plantio e doses de fertilizantes;
- Treinar e comparar múltiplos algoritmos (Rede Neural Artificial, Regressão Linear, SVR e Floresta Aleatória);
- Oferecer uma interface gráfica funcional, tornando o uso dos modelos acessível a agrônomos, consultores e produtores;
- Gerar previsões aplicáveis a cenários reais, como estudo de caso com diferentes doses de adubação;
- Transformar modelos acadêmicos em ferramenta prática, ocupando uma lacuna existente nos estudos analisados.

Assim, o Guiafertil avança além da literatura ao propor um sistema completo, voltado ao uso prático no setor produtivo, incorporando técnicas de IA em um ambiente integrado e orientado à tomada de decisão no campo.

3 MATERIAL E MÉTODOS

O Desafio Nacional de Máxima Produtividade de Soja, promovido anualmente pelo Comitê Estratégico Soja Brasil (CESB), tem como objetivo reconhecer os produtores com maior desempenho produtivo nos sistemas irrigado e de sequeiro. Nesse contexto, equipes técnicas especializadas realizam auditorias em lavouras distribuídas por todo o território nacional, assegurando a confiabilidade dos dados e promovendo a disseminação de conhecimentos e práticas agronômicas entre produtores e consultores. Considerando o rigor técnico e a qualidade dessas informações, a base de dados utilizada neste estudo foi obtida a partir dos "Cases Campeões" do CESB, reunindo registros de lavouras de alta produtividade empregados na construção e avaliação dos modelos preditivos desenvolvidos (CESB, s.d.).

A metodologia adotada neste estudo foi estruturada para integrar dados agronômicos provenientes de lavouras auditadas pelo Comitê Estratégico Soja Brasil (CESB), aplicar técnicas de pré-processamento, treinar modelos de aprendizado de máquina supervisionado e, por fim, disponibilizar esses resultados em um software funcional denominado Guiafertil.

O processo metodológico adotado neste trabalho é composto por cinco etapas principais, estruturadas de maneira sequencial e interdependente, garantindo rigor e coerência na construção do modelo preditivo. A primeira etapa consiste na organização e caracterização da base de dados, momento em que se verifica a integridade, a consistência e a relevância das informações coletadas. Essa fase é fundamental para compreender a estrutura dos dados, identificar possíveis limitações e assegurar que o conjunto esteja adequado para as análises subsequentes.

Na segunda etapa, desenvolve-se a arquitetura do software, definindo sua lógica interna, os módulos que o compõem, as funcionalidades que serão disponibilizadas ao usuário e o fluxo de informações entre as diferentes partes do sistema. Essa definição estrutural é essencial para garantir que o software seja funcional, eficiente e capaz de integrar corretamente todas as etapas do processo preditivo.

A terceira etapa corresponde ao pré-processamento das variáveis, que envolve a preparação técnica dos dados para sua utilização nos modelos de aprendizado de máquina. Nessa fase, são realizados procedimentos como tratamento de valores ausentes, transformações e normalizações, além da codificação de variáveis categóricas. Essa etapa busca tornar os dados compatíveis com as exigências dos algoritmos, minimizando ruídos e garantindo uma base adequada para o treinamento do modelo.

Na sequência, ocorre a configuração e o treinamento dos modelos de predição. Aqui, são definidos e ajustados os hiperparâmetros, selecionados os algoritmos mais adequados ao problema e executado o processo de aprendizado, no qual o modelo identifica padrões presentes nos dados. O objetivo é encontrar uma estrutura preditiva capaz de gerar estimativas consistentes e alinhadas com o comportamento real observado.

Por fim, a quinta etapa consiste na avaliação do desempenho dos modelos. Nessa fase, analisam-se métricas de precisão e confiabilidade, verificando o grau de acurácia das previsões geradas. Essa avaliação permite validar o modelo, identificar possíveis melhorias e garantir a qualidade dos resultados obtidos. Assim, as cinco etapas metodológicas se complementam, formando um processo robusto e sistematizado para o desenvolvimento das predições.

A Figura 5 apresenta o fluxograma geral da arquitetura funcional do Guiafertil, destacando o fluxo sequencial: entrada dos dados, pré-processamento, modelagem e avaliação.

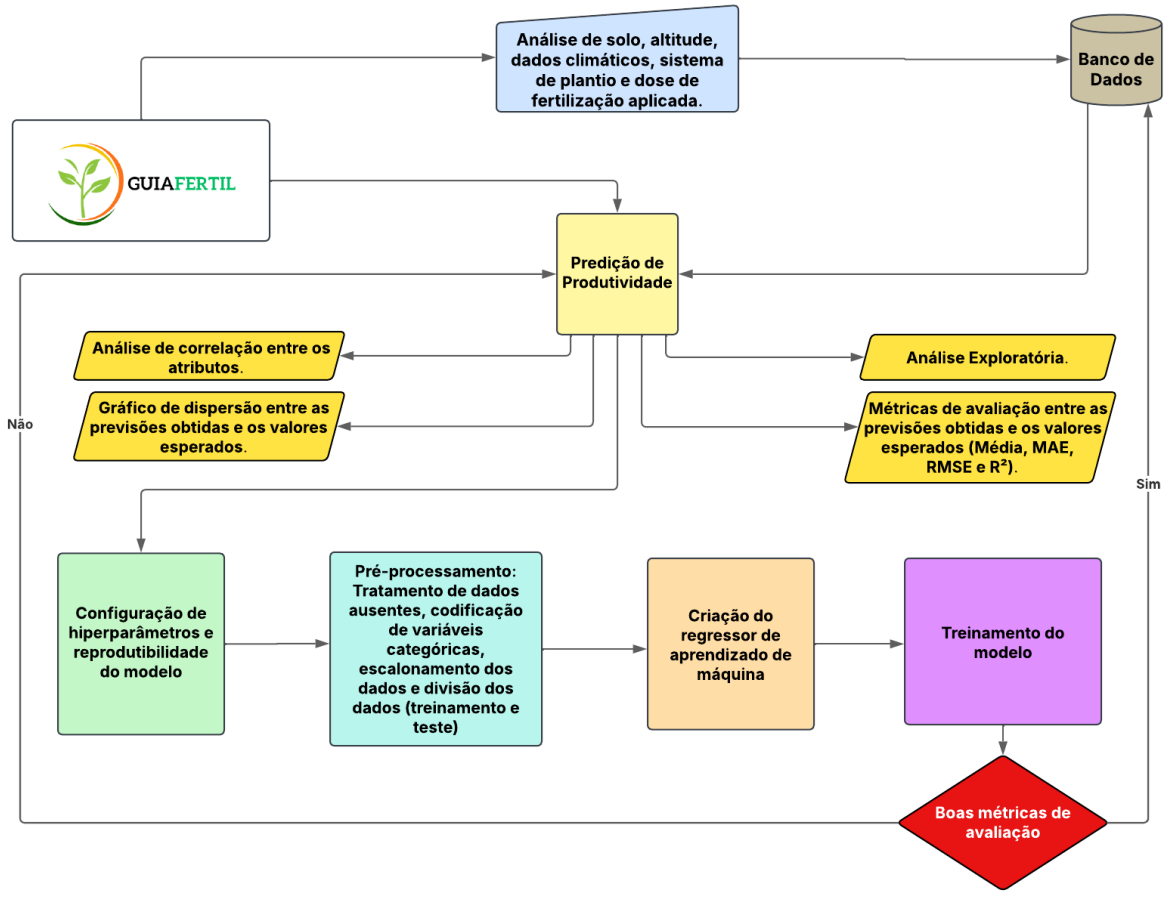


FIGURA 5 . Arquitetura funcional do software Guiafertil, destacando o fluxo de entrada dos dados, pré-processamento, modelagem e avaliação. Fonte: Elaborado pelo autor.

3.1 BASE DE DADOS

A base utilizada neste estudo extraída integralmente dos Cases Campeões do CESB, contém 37 registros, compostos por atributos relacionados ao solo, clima, altitude, sistema de plantio, fertilização aplicada e produtividade final. Esses dados foram organizados em matriz onde cada linha representa uma lavoura auditada e cada coluna corresponde a um atributo.

Para fins de modelagem, a divisão da base seguiu o padrão estabelecido na literatura:

- 80% dos dados destinados ao treinamento dos modelos;
- 20% dos dados reservados para avaliação de desempenho.

Essa proporção foi escolhida para garantir variabilidade suficiente no conjunto de treinamento, ao mesmo tempo em que mantém um subconjunto adequado para medir a capacidade de generalização dos modelos.

3.2 AMBIENTE DO SOFTWARE GUIAFERTIL

O ambiente do Guiafertil foi desenvolvido em Python, explorando recursos modernos que conferem robustez, flexibilidade e alto desempenho ao sistema. A solução integra o banco de dados SQLite3, que se destaca pela leveza, portabilidade e eficiência no armazenamento local de informações, permitindo operações rápidas mesmo em ambientes com recursos limitados. Para a construção da interface gráfica, foi empregada a biblioteca PySide6, que possibilita o desenvolvimento de telas intuitivas, responsivas e visualmente organizadas, proporcionando uma experiência de uso amigável e voltada às necessidades do gerenciamento de informações agrícolas. A combinação dessas tecnologias oferece um ecossistema integrado, capaz de unir simplicidade operacional, alta capacidade de processamento e facilidade de expansão, características fundamentais para aplicações que envolvem análise de dados e predição de produtividade.

O software é composto por 4 módulos principais: Cadastro de Usuários, Cadastro de Produtores, Cadastro de Atributos (dados de análises de solo, informações climáticas, adubação aplicada, média de produtividade alcançada, análise exploratória da produtividade) e o Ambiente de Predição de Produtividade que utiliza uma variedade de algoritmos de aprendizado supervisionado para predições como: Redes Neurais, Regressão Linear, SVR (Regressão de Vetores de Suporte) e Floresta Aleatória.

A Figura 6 apresenta a interface de cadastro de dados agronômicos do software Guiafertil. Nessa tela, o usuário pode inserir informações relacionadas às análises de solo, condições climáticas, produtividade da lavoura e também ao tipo de adubação utilizada. Entre os principais campos disponíveis, destacam-se os atributos químicos do solo, como pH e teores de nutrientes (P, K, Ca, Mg, entre outros), além de variáveis relacionadas ao ambiente de cultivo, como sistema de plantio, safra, cultura e manejo nutricional. A interface foi projetada para facilitar a organização e o armazenamento dos dados, permitindo o registro estruturado de informações de diferentes talhões. Esses dados são posteriormente utilizados na análise exploratória e no treinamento dos modelos de aprendizado de máquina, sendo fundamentais para a geração das predições de produtividade.

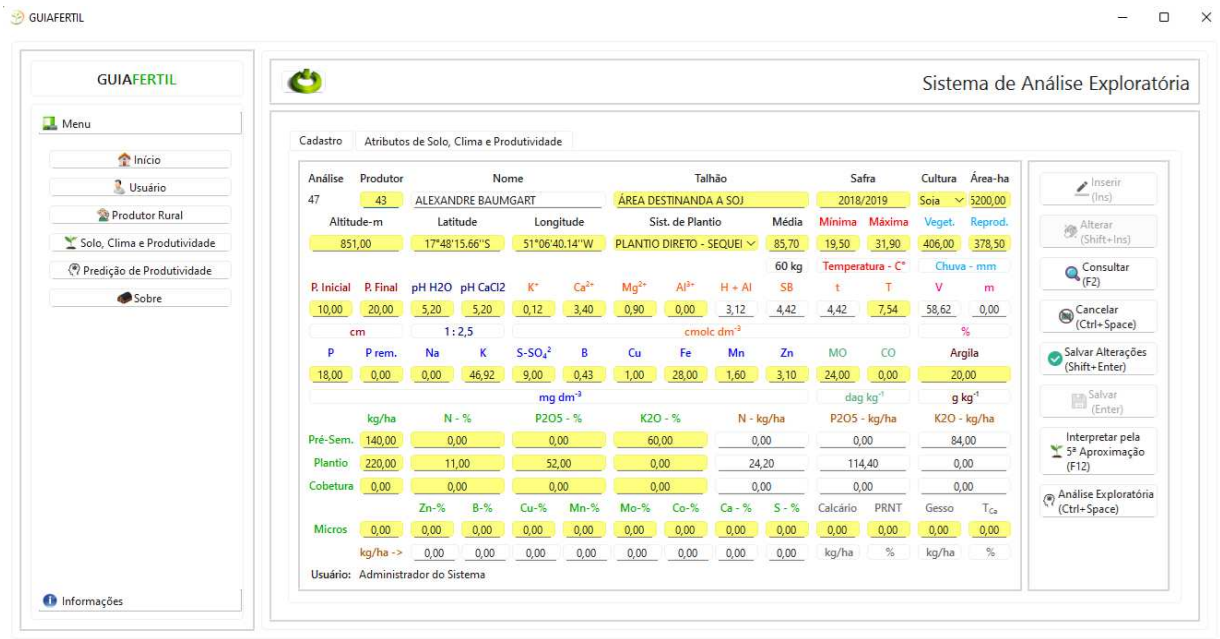


FIGURA 6 . Interface de cadastro utilizada para inserir os dados agronômicos no sistema Guiafertil. Fonte: Elaborado pelo autor.

A Figura 7 apresenta o ambiente de predição de produtividade do software Guiafertil durante a etapa de treinamento dos modelos de aprendizado de máquina. Nessa interface, o usuário pode configurar os principais parâmetros utilizados no processo de treinamento. Além disso, a tela permite a seleção dos algoritmos a serem utilizados, incluindo Rede Neural, Regressão Linear, SVR (Regressão de Vetores de Suporte) e Floresta Aleatória. Também é possível definir parâmetros específicos de cada modelo, como o número de épocas (epochs),

tamanho do lote (batch size), taxa de aprendizado (learning rate) e número de repetições (folds) da rede neural, o kernel da SVR e a quantidade de árvores da Floresta Aleatória. A interface apresenta ainda opções de validação cruzada, controle de aleatoriedade por meio de sementes (seeds) e armazenamento dos modelos treinados. Esses recursos permitem ao usuário ajustar e avaliar diferentes configurações, sendo fundamentais para a obtenção de modelos preditivos mais precisos e confiáveis.

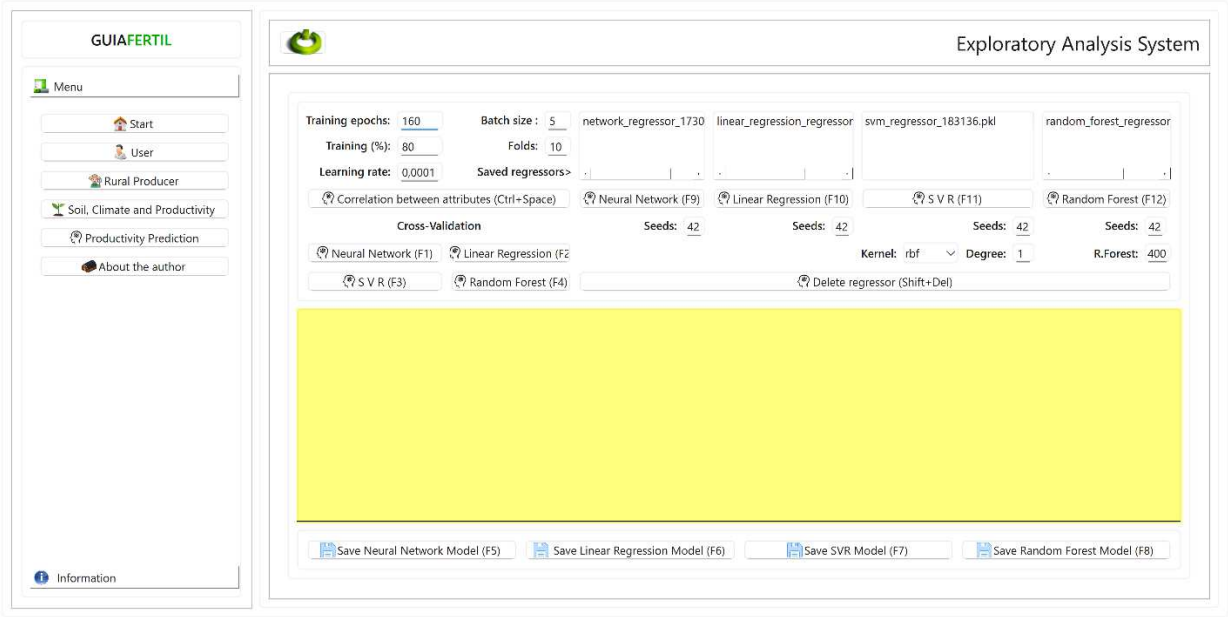


FIGURA 7 . Ambiente de aprendizado de máquina do Guiafertil, utilizado para configurar e treinar os modelos de predição de produtividade. Fonte: Elaborado pelo autor.

3.3 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento de dados é uma etapa fundamental em projetos de aprendizado de máquina, com o objetivo principal de melhorar a qualidade dos dados e, consequentemente, o desempenho do modelo proposto.

As etapas realizadas neste projeto durante o pré-processamento foram :

- Tratamento de dados faltantes: Valores ausentes foram tratados conforme a natureza da variável, assegurando consistência na matriz de dados utilizada pelos algoritmos.
- Codificação de variáveis categóricas: Variáveis como sistema de plantio foram transformadas através de OneHotEncoding, permitindo sua utilização em algoritmos matemáticos que exigem valores numéricos.

- Escalonamento dos dados: A normalização foi aplicada para ajustar a escala das variáveis numéricas, evitando que atributos com magnitudes elevadas dominassem o processo de treinamento.
- Controle da aleatoriedade: Para assegurar reprodutibilidade, foi estabelecida uma semente global (*random seed*) tanto no Python quanto no NumPy, além da habilitação de operações determinísticas nos modelos.
- Divisão treino–teste: Foi empregado o método train_test_split da biblioteca scikit-learn, resultando na separação dos dados em 80% para treinamento e 20% para teste.

Outro aspecto essencial antes de iniciar o treinamento dos modelos de aprendizado de máquina é a possibilidade de configurar a aleatoriedade global no Python e no NumPy, além de definir a semente aleatória para assegurar a reprodutibilidade do treinamento.

3.4 REDE NEURAL

Grus (2021) destaca a Rede Neural Artificial como um modelo preditivo baseado na dinâmica do cérebro, que tem uma série de neurônios conectados. Cada um deles analisa as saídas dos outros neurônios ligados nele, faz um cálculo e, em seguida, irá disparar (se o valor calculado exceder um limite) ou não disparar (se não exceder um limite).

Neste projeto foi usando a biblioteca TensorFlow e Keras que cria um modelo de Rede Neural sequencial que define sua arquitetura, especifica o otimizador, a função de perda e as métricas de avaliação (RIBEIRO, 2021).

A função de perda no treinamento de modelos de aprendizado de máquina, é usada para quantificar o erro entre as previsões do modelo e os valores reais. Essa função guia o processo de aprendizado, ajudando o algoritmo de otimização a ajustar os pesos do modelo para minimizar os erros.

A métrica de desempenho, após o modelo ser treinado, é usada para avaliar seu desempenho em um conjunto de dados de validação ou teste. Nesse contexto, ele não está diretamente envolvido no ajuste dos pesos do modelo, mas sim em medir quão bem o modelo aprendeu a prever os valores desejados.

O otimizador Adam é o algoritmo responsável no estudo por ajustar os pesos da Rede Neural durante o treinamento para minimizar a função de perda e o método train_test_split da biblioteca scikit-learn para dividir conjunto de dados em dois subconjuntos: um destinado ao treinamento do modelo e outro para avaliação (SALAM, 2024).

Os dados extraídos dos cases de sucesso do CESB foi dividido em dois subconjuntos: um para treinamento do modelo, configurado para 80%, e outro para avaliação, configurado para 20%.

Os hiperparâmetros utilizados para o treinamento do modelo de Rede Neural estão descritos a seguir:

- epochs: 160;
- batch_size: 5;
- learning_rate: 0,0001;
- kernel_initializer: 42 (valor da semente);

Os demais parâmetros que não foram configurados e adotaram os valores padrões de suas respectivas classes.

O processo de definição da arquitetura e dos hiperparâmetros não foi arbitrário. Foram realizados múltiplos testes utilizando validação cruzada, variando número de neurônios, quantidade de camadas, taxa de aprendizado e demais parâmetros do otimizador.

Os valores selecionados 160 epochs, batch_size igual a 5, learning rate ajustado e seed fixa demonstraram melhor desempenho médio, apresentando menor erro absoluto e maior estabilidade durante os testes de validação cruzada.

Assim, os hiperparâmetros aqui apresentados correspondem à configuração ideal identificada experimentalmente para este conjunto de dados.

O código-fonte completo utilizado na implementação da Rede Neural encontra-se disponível no repositório oficial do projeto no GitHub no link https://github.com/dariosilvamelo/guiafertil/blob/main/neural_network_functions.py.

3.5 REGRESSÃO LINEAR

Mueller (2020) define a Regressão Linear como uma reta que passa por uma série de coordenadas x/y, determinando a localização de um ponto de dados. Os pontos de dados nem sempre estão diretamente na reta, mas ela mostra onde estariam, em um mundo perfeito de coordenadas lineares. Com a reta, é possível predizer o valor de y, tendo o valor de x determinado. Quando há apenas uma variável preditora, temos uma Regressão Linear Simples e quando há muitas, temos uma Regressão Linear Múltipla.

A classe LinearRegression do módulo sklearn.linear_model é uma implementação de regressão linear Simples e múltipla, utilizada para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes através de uma equação linear neste estudo (PATIL, 2023).

Os hiperparâmetros utilizados para o treinamento do modelo de Regressão Linear estão descritos a seguir:

- fit_intercept: True;
- normalize: False;
- copy_X: True;
- n_jobs: None;
- random_state: 42 (valor da semente);

Os demais parâmetros que não foram configurados e adotaram os valores padrões de suas respectivas classes.

Os valores dos hiperparâmetros foram definidos após a realização de testes sistemáticos com validação cruzada, nos quais diferentes combinações de opções foram avaliadas. O conjunto acima apresentou o menor erro médio entre as alternativas testadas, tornando-se a configuração mais adequada para o conjunto de dados utilizado.

O código-fonte completo utilizado na implementação da Regressão Linear encontra-se disponível no repositório oficial do projeto no GitHub no link https://github.com/dariosilvamelo/guiafertil/blob/main/regression.py.

3.6 SVM

SVM, (Máquina de Vetores de Suporte), é uma técnica de aprendizado de máquina amplamente aplicada em tarefas de classificação e regressão. Géron (2021) explica que a Regressão por Máquinas de Vetores de Suporte, é também conhecida como SVR (Regressão de Vetores de Suporte), utilizada para modelar relações entre variáveis em cenários de predição. Diferentemente da Regressão Linear tradicional, a SVR busca não apenas ajustar uma linha que represente os dados, mas também introduz o conceito de margem de tolerância (ε), que define um intervalo ao redor da linha de regressão dentro do qual os desvios não são penalizados. Essa margem define o intervalo em que os desvios entre os valores observados e os valores preditos são considerados aceitáveis, sem impacto na função objetivo do modelo. Aqueles que se

encontram dentro da margem de tolerância são tratados como suficientemente próximos à predição e, portanto, não influenciam na otimização do modelo. Em contraste, os pontos que estão fora dessa margem de tolerância são chamados de vetores de suporte, pois desempenham um papel fundamental na definição da linha de regressão. Aqueles pontos dentro da margem de tolerância (ε) não influenciam diretamente o ajuste do modelo, enquanto os pontos fora da margem impactam a construção da curva, forçando o modelo a levar em conta os desvios mais significativos. A classe SVR utilizada no projeto é da biblioteca sklearn.svm que implementa o SVR, que é a versão para regressão do algoritmo SVM (YERRABOLU, 2024).

Os hiperparâmetros utilizados para o treinamento do modelo SVR estão descritos a seguir:

- kernel: 'rbf';
- C: 1.0;
- epsilon: 0.1;
- degree: 3;
- random_state: 42 (valor da semente);

Os demais parâmetros que não foram configurados e adotaram os valores padrões de suas respectivas classes.

Para definição desses valores, foram realizados vários experimentos combinando diferentes valores de C, epsilon, kernel e degree, utilizando validação cruzada para medir o desempenho de cada configuração. Os hiperparâmetros apresentados foram aqueles que demonstraram o melhor equilíbrio entre erro preditivo e estabilidade do modelo, apresentando o desempenho mais consistente nos folds de validação.

O código-fonte completo utilizado na implementação do SVM encontra-se disponível no repositório oficial do projeto no GitHub no link https://github.com/dariosilvamelo/guiafertil/blob/main/support_vectors.py.

3.7 FLORESTA ALEATÓRIA

Géron (2021) explica que as respostas agregadas geralmente superam a resposta de um único especialista, fenômeno conhecido como "sabedoria das multidões". Nesse contexto, um grupo de preditores tende a produzir resultados melhores do que o melhor preditor individual. Esse conjunto de predições é denominado ensemble (agrupamento), e a técnica de aprendizado

que utiliza esse conceito é conhecida como método ensemble. A Floresta Aleatória é um algoritmo de aprendizado de máquina baseado em um conjunto de árvores de decisão. Ele combina múltiplas árvores para melhorar a precisão e a generalização do modelo. Esse método faz parte da categoria dos métodos ensemble e pode ser usado tanto para classificação quanto para regressão.

A classe RandomForestRegressor utilizada neste projeto da biblioteca scikit-learn é uma implementação do algoritmo de Floresta Aleatória voltada para problemas de regressão. Essa técnica utiliza um conjunto de árvores de decisão para estimar um valor contínuo, combinando os resultados de múltiplas árvores para produzir previsões mais robustas e precisas (HALIM, 2023).

Os hiperparâmetros utilizados para o treinamento do modelo Floresta Aleatória estão descritos a seguir:

- n_estimators: 400;
- random_state: 42 (valor da semente);
- bootstrap: True;

Os demais parâmetros que não foram configurados e adotaram os valores padrões de suas respectivas classes.

Durante o processo de modelagem, foram conduzidos testes sucessivos com validação cruzada, variando o número de árvores, profundidade máxima, critérios de divisão e amostragem. O valor de 400 árvores apresentou o melhor desempenho médio, reduzindo o erro absoluto sem causar sobre ajuste.

Assim, os hiperparâmetros selecionados representam a melhor configuração obtida empiricamente para o conjunto de dados analisado.

O código-fonte completo utilizado na implementação da Floresta Aleatória encontra-se disponível no repositório oficial do projeto no GitHub no link https://github.com/dariosilvamelo/guiafertil/blob/main/random_forest.py.

4 RESULTADOS E DISCUSSÃO

Este capítulo tem como objetivo apresentar, analisar e interpretar os resultados obtidos pelos modelos de aprendizado de máquina aplicados à base de dados dos Cases Campeões (CESB), bem como discutir suas implicações práticas no manejo da cultura da soja. Inicialmente, são apresentados os valores previstos por cada modelo e suas respectivas métricas de desempenho, permitindo comparar a acurácia entre as diferentes abordagens. Em seguida, são analisados os gráficos de dispersão das predições, destacando padrões, tendências e limitações observadas. Posteriormente, é realizado um estudo de caso para avaliar o comportamento dos modelos em cenários de manejo. Por fim, este capítulo apresenta a avaliação qualitativa do software Guiafertil realizada por agrônomos, discutindo a utilidade, aplicabilidade e potencial de uso da ferramenta no contexto agrícola.

A Tabela 3 apresenta, de forma detalhada, os resultados individuais obtidos pelos quatro modelos de aprendizado de máquina para o conjunto de teste, incluindo os valores reais de produtividade (yi), as previsões geradas (ŷi), o erro absoluto ($|y_i - \hat{y}_i|$), o termo $(|y_i - \hat{y}_i| - MAE)^2$ utilizado no cálculo da variabilidade dos erros, o erro quadrático $(y_i - \hat{y}_i)^2$, além da variância dos dados $(y_i - \bar{y}_i)^2$, que representa o afastamento de cada valor real em relação à média da produtividade. Esses valores constituem a base necessária para o cálculo das métricas síntese apresentadas posteriormente.

TABELA 3 . Resultados das previsões dos modelos de aprendizado de máquina: valores reais, valores previstos, erro absoluto, erro quadrático e a variância dos dados. Fonte: Elaborado pelo autor.

Rede Neural (MLP)					
Produtividade Real (y_i)	Previsão ($\hat{y}_i$)	Erro Absoluto ($\|y_i - \hat{y}_i\|$)	(Erro Absoluto - MAE)²	Erro quadrático $(y_i - \hat{y}_i)^2$	Variância dos Dados $(y_i - \bar{y}_i)^2$
90,0000	91,0438	1,0438	9,0522	1,0895	65,6100
70,0000	77,4858	7,4858	11,7880	56,0377	141,6100
92,0000	86,4027	5,5974	2,3867	31,3303	102,0100
73,5000	77,1419	3,6419	0,1686	13,2634	70,5600
82,0000	75,6923	6,3077	5,0862	39,7875	0,0100
80,0000	85,1857	5,1857	1,2841	26,8911	3,6100
82,0000	84,3755	2,3755	2,8123	5,6429	0,0100
85,7000	86,4820	0,7820	10,6959	0,6115	14,4400
655,2000	**663,8096**	**32,4197**	**43,2740**	**174,6538**	**397,8600**

Regressão Linear					
Produtividade Real (y_i)	Previsão ($\hat{y}_i$)	Erro Absoluto ($\lvert y_i - \hat{y}_i \rvert$)	(Erro Absoluto - MAE)²	Erro quadrático $(y_i - \hat{y}_i)^2$	Variância dos Dados $(y_i - \bar{y}_i)^2$
90,0000	97,3934	7,3934	0,1440	54,6624	65,6100
70,0000	85,9367	15,9367	66,6487	253,9788	141,6100
92,0000	97,6273	5,6273	4,6033	31,6666	102,0100
73,5000	70,8004	2,6996	25,7374	7,2881	70,5600
82,0000	70,6779	11,3221	12,5973	128,1903	0,0100
80,0000	69,0353	10,9647	10,1876	120,2236	3,6100
82,0000	75,8950	6,1050	2,7818	37,2708	0,0100
85,7000	87,8340	2,1340	31,7969	4,5538	14,4400
655,2000	**655,2000**	**62,1828**	**154,4970**	**637,8345**	**397,8600**
SVM					
Produtividade Real (y_i)	Previsão ($\hat{y}_i$)	Erro Absoluto ($\lvert y_i - \hat{y}_i \rvert$)	(Erro Absoluto - MAE)²	Erro quadrático $(y_i - \hat{y}_i)^2$	Variância dos Dados $(y_i - \bar{y}_i)^2$
90,0000	86,5726	3,4274	1,2680	11,7470	65,6100
70,0000	81,2404	11,2404	44,7155	126,3471	141,6100
92,0000	84,6273	7,3727	7,9484	54,3574	102,0100
73,5000	78,8524	5,3524	0,6383	28,6479	70,5600
82,0000	78,8515	3,1485	1,9739	9,9131	0,0100
80,0000	82,3288	2,3288	4,9492	5,4231	3,6100
82,0000	83,5224	1,5224	9,1874	2,3176	0,0100
85,7000	83,6650	2,0350	6,3424	4,1414	14,4400
655,2000	**659,6602**	**36,4276**	**77,0231**	**242,8946**	**397,8600**
Floresta Aleatória					
Produtividade Real (y_i)	Previsão ($\hat{y}_i$)	Erro Absoluto ($\lvert y_i - \hat{y}_i \rvert$)	(Erro Absoluto - MAE)²	Erro quadrático $(y_i - \hat{y}_i)^2$	Variância dos Dados $(y_i - \bar{y}_i)^2$
90,0000	85,1409	4,8591	0,1253	23,6104	65,6100
70,0000	77,3101	7,3101	7,8678	53,4376	141,6100
92,0000	82,7200	9,2800	22,7994	86,1187	102,0100
73,5000	78,5540	5,0540	0,3012	25,5427	70,5600
82,0000	76,4284	5,5716	1,1374	31,0430	0,0100
80,0000	81,2105	1,2105	10,8545	1,4654	3,6100
82,0000	83,7821	1,7821	7,4150	3,1759	0,0100
85,7000	84,7263	0,9737	12,4707	0,9482	14,4400
655,2000	**649,8723**	**36,0411**	**62,9712**	**225,3417**	**397,8600**

Os resultados mostram que a Rede Neural apresentou erros absolutos menores e mais consistentes ao longo das oito observações do conjunto de teste. O modelo registrou valores de erro absoluto predominantemente baixos, o que demonstra capacidade de aproximação das produtividades reais. Além disso, o termo $(|y_i – \hat{y}_i| - MAE)^2$ indica menor dispersão em relação ao MAE, refletindo maior estabilidade preditiva. Os valores de erro quadrático também se mantiveram baixos, contribuindo para um RMSE reduzido, indicando que a Rede Neural penaliza menos erros extremos quando comparada aos demais modelos.

O modelo de Floresta Aleatória apresentou desempenho semelhante ao da Rede Neural, com erros absolutos relativamente baixos e distribuição consistente. Apesar de apresentar leves oscilações, os valores de erro quadrático permaneceram dentro de limites aceitáveis, indicando boa capacidade de generalização. A variância dos dados permanece constante entre os modelos, pois é calculada apenas com base nas produtividades reais, reforçando que as diferenças nas métricas dependem exclusivamente da qualidade das previsões.

O modelo SVR apresentou desempenho intermediário. Embora tenha fornecido previsões adequadas em diversas observações, alguns erros individuais foram maiores que os registrados pelos modelos mais robustos. Isso pode ser observado nos termos de erro absoluto e erro quadrático, que exibem maior variabilidade. Ainda assim, o SVR produziu resultados razoáveis, com desempenho superior ao da Regressão Linear.

Por outro lado, o modelo de Regressão Linear apresentou o pior desempenho entre os avaliados. Os erros absolutos foram substancialmente maiores em várias observações, refletindo dificuldade em capturar relações não lineares existentes nos dados agronômicos. Os valores elevados de erro quadrático demonstram que o modelo comete erros extremos com maior frequência, o que se reflete diretamente no alto valor de RMSE. A dispersão dos erros também foi elevada, evidenciada pelo $(|y_i – \hat{y}_i| - MAE)^2$, indicando baixa consistência preditiva.

A Tabela 4 apresenta as métricas de desempenho dos modelos de aprendizado máquina. A tabela sintetiza o desempenho dos modelos com base nas métricas MEAN das produtividades reais e previstas, MAE, RMSE e R^2, incluindo a apresentação de MAE acompanhado do desvio padrão (MEAN ± STD). Os resultados reforçam as observações da Tabela 3.

TABELA 4 . Métricas de desempenho dos modelos de aprendizado de máquina após o treinamento. Fonte: Elaborado pelo autor.

Modelos	**MEAN (Produtividade Real) sacas/ha**	**MEAN (Previsões) sacas/ha**	**MAE sacas/ha**	**MAE (MEAN ± STD) sacas/ha**	**RMSE sacas/ha**	**R² sacas/ha**
Rede Neural (MLP)	81,90	82,98	4,05	4,05 ± 2.49	4,67	0,56
Regressão Linear	81,90	81,90	7,77	7.77 ± 4.70	8,93	-0,60
SVR	81,90	82,46	4,55	4.55 ± 3.32	5,51	0,39
Floresta Aleatória	81,90	81,23	4,51	4.51 ± 3.00	5,31	0,43

A Rede Neural apresentou o menor MAE (4,05) e o maior coeficiente de determinação ($R^2 = 0,56$), evidenciando melhor ajuste global ao conjunto de dados. A presença de um desvio padrão moderado (±2,49) confirma estabilidade nas previsões. Na sequência, a Floresta Aleatória apresentou MAE de 4,51 e R^2 de 0,43, demonstrando bom desempenho e proximidade com a Rede Neural.

O SVR apresentou MAE de 4,55 e R^2 de 0,39, desempenho inferior mas ainda aceitável. O modelo de Regressão Linear novamente destacou-se negativamente, com MAE elevado (7,77), RMSE de 8,93 e valor de R^2 negativo (-0,60). Um coeficiente de determinação negativo indica que o modelo performou pior do que uma predição simples baseada na média das produtividades reais, reforçando que uma relação linear não é capaz de representar adequadamente o comportamento da produtividade da soja diante das variáveis agronômicas utilizadas.

De forma geral, a análise conjunta das Tabelas 3 e 4 demonstra que os modelos não lineares especialmente a Rede Neural e a Floresta Aleatória foram os mais eficazes para a tarefa de estimação de produtividade. Esses modelos apresentaram erros menores, maior estabilidade e maior capacidade de explicação da variabilidade dos dados, evidenciando que a produtividade da soja possui comportamento complexo e não linear, justificando o uso de algoritmos mais avançados de aprendizado de máquina.

É fundamental buscar constantemente o aprimoramento da acurácia das previsões, expandindo a base de dados e ajustando os hiperparâmetros à medida que ela cresce, assegurando uma maior capacidade preditiva para aplicações na agricultura de precisão.

A Tabela 5 a seguir demonstra um estudo de caso prático.

TABELA 5 . Estudo de caso prático: dados agronômicos da lavoura, cenários de adubação e previsões dos modelos. Fonte: Elaborado pelo autor.

Dados de uma lavoura de soja							
Safra Agrícola	2022/2023						
Área (ha)	4060						
Altitude	478						
Latitude	-17.196388						
Longitude	-54.858611						
Sistema de Plantio	PLANTIO DIRETO - SEQUEIRO						
Temperatura Mínima (°C)	13						
Temperatura Máxima (°C)	37						
Chuva no Vegetativo (mm)	52,60						
Chuva no Reprodutivo (mm)	769,20						
pH CaCl2 (1 : 2,5)	5						
Argila (g/kg)	524,00						
MO (dag/kg)	41,00						
CO (dag/kg)	24,00						
K (cmolc/dm3)	0,27						
Ca (cmolc/dm3)	2,90						
Mg (cmolc/dm3)	1,10						
P (mg/dm3)	13,00						
K (mg/dm3)	105,57						
S-SO4 (mg/dm3)	9,00						
B (mg/dm3)	0,21						
Cu (mg/dm3)	1,00						
Fe (mg/dm3)	38,00						
Mn (mg/dm3)	0,80						
Zn (mg/dm3)	2,40						
Registro da base de dados	**Adubação formula 00-23-20 (kg/ha)**			**Produtividade Esperada (sacas/ha)**			**Produtividade Real (sacas/ha)**
				Rede Neural	**Regressão Linear**	**SVR**	**Floresta Aleatória**
	Quant. 350	**N**	0,00	83,68	81,06	82,78	81,05	83,70
		P	80,50					
		K	70,00					
Estudo de caso 01	**Quant.** 250	**N**	0,00	83,53	80,36	82,92	80,91	-
		P	57,50					
		K	50,00					
Estudo de caso 02	**Quant.** 450	**N**	0,00	83,69	81,76	82,54	80,88	-
		P	103,50					
		K	90,00					
Ajustar a adubação impacta diretamente os custos, sem grandes variações na produtividade.								

A Tabela 5 apresenta os dados agronômicos de uma lavoura de soja da safra 2022/2023, contemplando características ambientais, climáticas e químicas do solo, bem como informações referentes ao manejo da área. Entre as variáveis registradas estão altitude, localização geográfica, sistema de plantio, temperaturas mínima e máxima registradas no ciclo, precipitação ocorrida nos estágios vegetativo e reprodutivo, além de atributos químicos como pH, teores de argila, matéria orgânica, carbono orgânico, bem como concentrações de macro e micronutrientes. Esses dados constituem o conjunto de variáveis utilizado como entrada no software Guiafertil para estimativa de produtividade com os modelos de aprendizado de máquina.

Com base nesse registro, foi avaliado inicialmente o cenário da adubação real aplicada na área, utilizando a fórmula 00–23–20 na dose de 350 kg ha^{-1}. Para esse manejo, os modelos apresentaram produtividades previstas entre 81,05 e 83,68 sacas/ha, sendo que a produtividade real da lavoura foi 83,70 sacas/ha. A proximidade entre os valores previstos e o valor observado demonstra que os modelos conseguem representar adequadamente o comportamento produtivo da área em manejo real.

No Estudo de Caso 01, simulou-se uma redução da adubação para 250 kg ha^{-1}, mantendo a mesma fórmula. Os modelos estimaram produtividades variando entre 80,36 e 82,92 sacas/ha, valores muito próximos aos obtidos no cenário original. Isso indica que a redução da adubação neste solo, dentro dessa faixa, não compromete a produtividade, sugerindo que o agricultor pode diminuir a dose de fertilizantes sem prejuízo ao rendimento, favorecendo a redução dos custos de produção.

No Estudo de Caso 02, simulou-se o cenário oposto, aumentando-se a adubação para 450 kg ha^{-1}. Novamente, os modelos apresentaram estimativas semelhantes às dos cenários anterior e original, com produtividades entre 80,88 e 82,54 sacas/ha. Esse comportamento demonstra que o aumento da adubação neste solo, não resultou em incremento produtivo, indicando que, nas condições químicas e climáticas da lavoura analisada, o sistema encontra-se próximo ao seu limite de resposta à adubação. Assim, quantidades superiores de fertilizante não se traduzem em ganho de produtividade.

Esses resultados reforçam a utilidade do software Guiafertil como ferramenta de apoio à tomada de decisão, permitindo ao produtor simular diferentes cenários de manejo e otimizar o uso de fertilizantes, alcançando melhor eficiência econômica sem comprometer o potencial produtivo.

Os gráficos de dispersão apresentados nas Figuras 8, 9, 10 e 11 oferecem uma representação ilustrativa da relação entre os valores reais de produtividade e as previsões geradas pelos modelos. Os gráficos de dispersão ilustram a relação entre os valores reais de produtividade (no eixo X) e os valores previstos pelos modelos (no eixo Y). Os pontos azuis representam as combinações entre os valores reais e as respectivas previsões. A linha vermelha tracejada indica o alinhamento ideal, onde os valores reais seriam exatamente iguais aos valores previstos (perfeita correspondência).

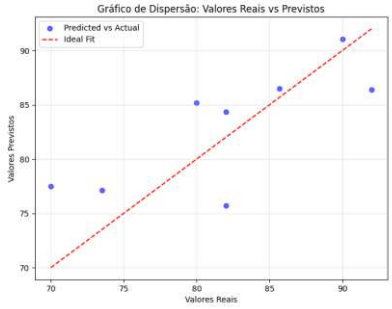
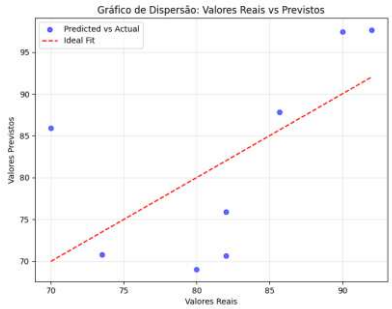


FIGURA 8 . Gráfico de dispersão da Rede Neural, comparando os valores reais de produtividade com os valores previstos pelo modelo. Fonte: Elaborado pelo autor.



.FIGURA 9 . Gráfico de dispersão da Regressão Linear, comparando os valores reais de produtividade com os valores previstos pelo modelo. Fonte: Elaborado pelo autor.

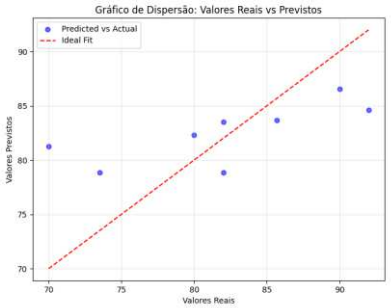


FIGURA 10 . Gráfico de dispersão do modelo SVM (SVR), comparando os valores reais de produtividade com os valores previstos pelo modelo. Fonte: Elaborado pelo autor.

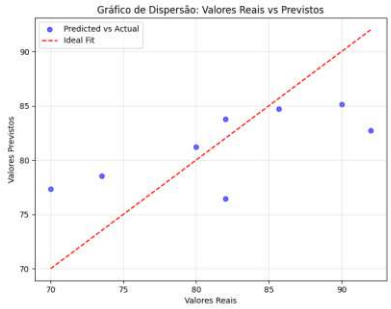


FIGURA 11 . Gráfico de dispersão do modelo Floresta Aleatória, comparando os valores reais de produtividade com os valores previstos pelo modelo. Fonte: Elaborado pelo autor.

A Tabela 6 mostra o resumo dos resultados comparativos entre os gráficos de dispersão.

TABELA 6 . Comparação visual do desempenho dos modelos com base nos gráficos de dispersão. Fonte: Elaborado pelo autor.

Modelo	**Distribuição dos Pontos no Gráfico**	**Precisão Geral**
Rede Neural	Pontos bem alinhados à linha ideal	Melhor modelo
Regressão Linear	Padrão disperso e sem relação clara	Pior modelo
SVR	Alguns pontos ajustados, outros dispersos	Desempenho intermediário
Floresta Aleatória	Boa aderência geral, mas com desvios	Desempenho intermediário

A Figura 12 apresenta o ambiente de analise exploratória, onde o usuário seleciona o modelo e visualiza a previsão de produtividade para um ambiente específico, com base nos atributos de solo, clima, sistema de plantio e adubação.

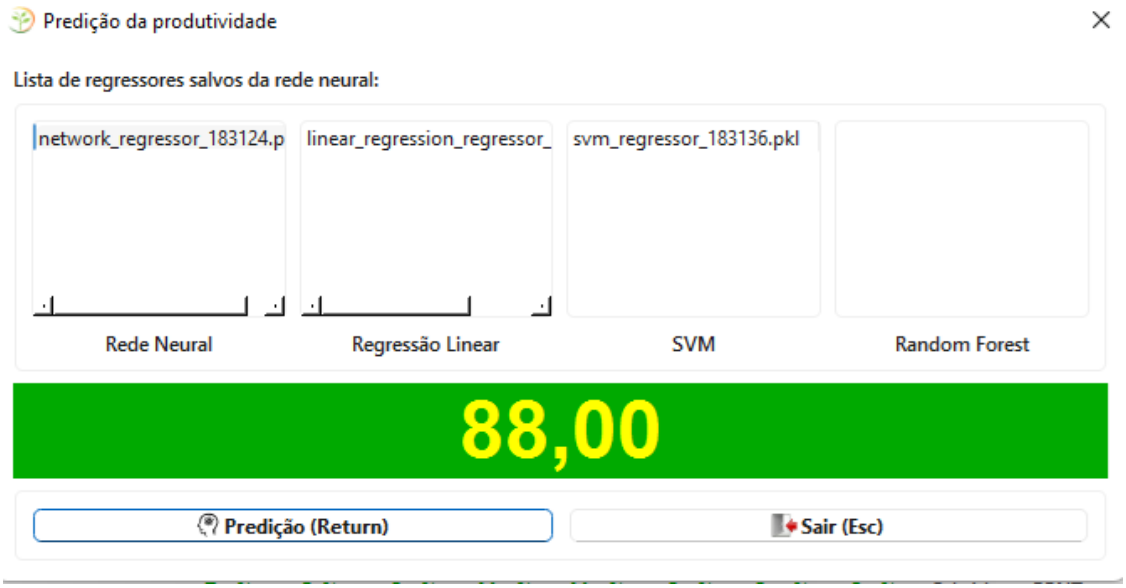


FIGURA 12 . Tela do ambiente de predição de produtividade do software Guiafertil, exibindo os modelos salvos e o valor previsto gerado pelo regressor selecionado. Fonte: Elaborado pelo autor.

Com o objetivo de avaliar a qualidade do software Guiafertil, suas funcionalidades foram apresentadas a cinco engenheiros agrônomos com experiência em manejo de culturas agrícolas e análise de dados agronômicos. Em seguida, foi aplicado um questionário para que eles pudessem avaliar a qualidade do sistema. A Figura 13 apresenta o questionário utilizado.

Questionário de Avaliação da Qualidade do Software "Guiafertil"

O entrevistado deverá responder ao questionário abaixo em uma escala qualitativa de 0 a 10, na qual 0 corresponde à pior avaliação possível e 10 à melhor. A justificativa da sua resposta é opcional.

Tópico	Subtópico/Módulo	N°	Questão	Resposta uma escala qualitativa de 0 a 10	Justificativa da Resposta (Opcional) Ex,: Questão n° xx : .....
Qualidade Geral do Software	Funcionalidade (Adequação Funcional)	01	O software atende às necessidades e requisitos esperados para sua função?		
	Desempenho e Eficiência	02	O tempo de resposta e o consumo de recursos do software são adequados?		
	Usabilidade	03	O software é fácil de aprender e utilizar?		
	Confiabilidade	04	O software opera sem falhas frequentes?		
	Eficiência em Uso	05	O software permite que os usuários realizem suas tarefas de forma eficiente?		
	Satisfação do Usuário	06	De modo geral, qual o seu nível de satisfação com o software?		
Avaliação por Módulo do Sistema	Cadastro de Usuários (Entrada)	07	O cadastro de usuários é fácil localizar e acessar o menu de cadastro?		
		08	A interface do cadastro de usuários permite adicionar/editar dados com facilidade?		
	Cadastro de Usuários (Saída)	09	O cadastro de usuários confirma o sucesso do cadastro de forma clara?		
		10	O login com as credenciais recém-criadas funciona corretamente?		
	Cadastro de Produtores (Entrada)	11	O cadastro de produtores é fácil localizar e acessar o menu de cadastro?		
		12	A interface do cadastro de produtores permite adicionar/editar dados com facilidade?		
	Cadastro de Produtores (Saída)	13	O cadastro de produtores confirma o sucesso do cadastro de forma clara?		
		14	A busca por produtores é eficaz e rápida?		
	Cadastro de Áreas e Dados Agronômicos (Entrada)	15	É possível associar corretamente as áreas cadastradas da lavoura aos produtores?		
		16	Os dados agronômicos são de fácil inserção?		
	Cadastro de Áreas e Dados Agronômicos (Saída)	17	Os dados no cadastro de áreas são exibidos corretamente no sistema?		
		18	A previsão de produtividade funciona com dados completos?		
	Ambiente de Predição de Produtividade (Entrada)	19	No ambiente de predição a seleção de dados para análise é clara e intuitiva?		
		20	No ambiente de predição é fácil configurar os parâmetros dos modelos?		
	Ambiente de Predição de Produtividade (Saída)	21	No ambiente de predição os gráficos e métricas são informativos?		
		22	No ambiente de predição o tempo de execução das análises é adequado?		
		23	O ambiente de predição permite exportar os modelos com facilidade?		
Avaliação Geral do Teste	Funcionamento Geral	24	Desempenho das funcionalidades foi eficiente sem erros e travamentos?		
		25	O sistema está pronto para uso em ambiente real?		

Monte Carmelo, _____ de _____ de 2025.

Assinatura do Avaliador : _____

Nome do Eng. Agrônomo : **CPF:**

FIGURA 13 . Questionário de avaliação da qualidade do software Guiafertil, elaborado para medir a percepção dos engenheiros agrônomos quanto à funcionalidade, eficiência, usabilidade e desempenho dos módulos do sistema. Fonte: elaborado pelo autor.

Os participantes foram selecionados por conveniência, considerando sua atuação profissional na área e familiaridade com processos de tomada de decisão no campo.

Após a apresentação do sistema, foi aplicado um questionário estruturado, elaborado com base em critérios de avaliação de software, contemplando aspectos como funcionalidade, eficiência, usabilidade e confiabilidade. As respostas foram registradas por meio de escala do tipo Likert, permitindo quantificar a percepção dos avaliadores em relação ao desempenho do sistema.

A Figura 13 apresenta o instrumento de avaliação utilizado, contendo as questões aplicadas aos especialistas.

Cabe destacar que, embora os resultados obtidos indiquem uma avaliação positiva do software, a amostra reduzida de participantes constitui uma limitação deste estudo. Dessa forma, como oportunidade para trabalhos futuros, sugere-se a ampliação do número de avaliadores, incluindo profissionais de diferentes regiões e perfis, a fim de obter uma análise mais abrangente e representativa da qualidade do sistema.

A Figura 14 apresenta o gráfico com os resultados obtidos por meio do questionário.

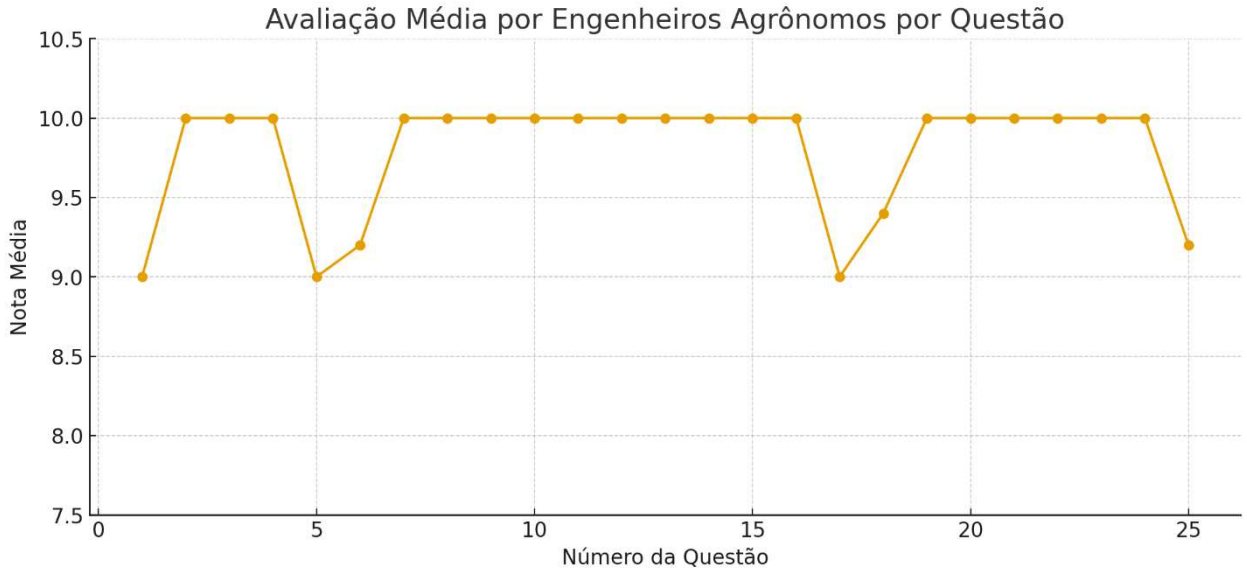


FIGURA 14 . Resultado do questionário de avaliação da qualidade do software Guiafertil, apresentando a média das notas atribuídas pelos engenheiros agrônomos para cada questão.
Fonte: Elaborado pelo autor.

O gráfico revela que, na opinião dos engenheiros agrônomos consultados, o Guiafertil demonstrou alta qualidade técnica e funcional, com pequenos pontos de atenção que podem ser ajustados para torná-lo ainda mais robusto e pronto para o uso no campo. Essas pequenas

variações são normais em fases de validação e podem guiar os próximos passos no refinamento do sistema.

O modelo do Guiafertil apresentado, pode abrir portas para explorar novas questões como:

- Análise preditiva em diferentes culturas agrícolas: Investigação da aplicabilidade do modelo a outras culturas além da soja, como milho e trigo.
- Impactos climáticos na produtividade agrícola: Integração de variáveis climáticas para prever como mudanças no clima podem impactar a produtividade.
- Otimização de recursos agrícolas: Estudos sobre como o modelo pode ser usado para otimizar o uso de insumos agrícolas.

A disseminação do software Guiafertil dentro e fora do grupo de usuários pode ser:

- No grupo-alvo:
  - Downloads e usuários ativos: Caso disponibilizado publicamente, é possível monitorar métricas como número de downloads e acessos para avaliar seu alcance.
  - Adoção em universidades e centros de pesquisa: A aceitação do software no meio acadêmico pode ser evidenciada por publicações que o citem como ferramenta utilizada.
- Fora do grupo-alvo:
  - Aplicações no setor comercial: Empresas agrícolas podem incorporar o modelo em plataformas destinadas à análise preditiva, ampliando seu uso em práticas empresariais.
  - Colaborações interdisciplinares: O software também pode ser explorado por profissionais de outras áreas, como economistas agrícolas e planejadores ambientais.
- Aplicações comerciais: Serviços de consultoria agrícola.
- Criação de spin-offs: Empresas podem surgir com base no software, oferecendo soluções como plataformas para análise de produtividade ou otimização de insumos agrícolas.

5 CONCLUSÕES

A proposta do presente trabalho atingiu seu objetivo ao desenvolver o Guiafertil, um software capaz de realizar análise exploratória de dados agronômicos e estimar o limiar de produtividade da soja utilizando algoritmos de inteligência artificial. A proposta central consistia em criar uma ferramenta que integrasse banco de dados, rotinas de pré-processamento, geração de gráficos, análise estatística e modelos preditivos, de modo a oferecer suporte técnico à tomada de decisão no manejo agrícola. Os resultados obtidos demonstram que essa integração foi alcançada de maneira eficaz, estruturando um sistema funcional e alinhado às demandas da agricultura de precisão.

Ao avaliar os resultados à luz dos objetivos específicos, verificou-se que a implementação dos modelos de aprendizado de máquina possibilitou identificar padrões relevantes entre atributos do solo, variáveis climáticas, altitude e sistema de plantio. Esses modelos, especialmente a Rede Neural Artificial e a Floresta Aleatória, apresentaram desempenho consistente, revelando que técnicas de IA são capazes de representar de forma eficiente a complexidade envolvida na produtividade agrícola. A etapa de pré-processamento e análise exploratória contribuiu significativamente para compreender a distribuição das variáveis e apoiar a seleção dos atributos mais influentes, reforçando a importância dessa fase no processo de modelagem.

A construção da interface gráfica do Guiafertil mostrou-se adequada, proporcionando um ambiente organizado em módulos de cadastro, gerenciamento de dados e predição. Essa estrutura favoreceu a usabilidade do sistema e permitiu ao usuário navegar pelas funcionalidades de maneira intuitiva e objetiva. A integração de métricas estatísticas e ferramentas de avaliação também contribuiu para dar transparência ao desempenho dos modelos, permitindo sua comparação e análise crítica.

A validação realizada por cinco engenheiros agrônomos ajudou a reforçar a qualidade técnica e funcional do sistema. As avaliações positivas recebidas indicam que o software atende às necessidades básicas para uso no campo e possui potencial para auxiliar profissionais em suas atividades de planejamento e recomendação de insumos. Esses resultados confirmam a hipótese de pesquisa, que defendia a capacidade das técnicas de IA em prever a produtividade da soja com maior precisão em comparação aos métodos convencionais utilizados atualmente.

Embora os resultados obtidos demonstrem o potencial do software Guiafertil como ferramenta de apoio à tomada de decisão agronômica, algumas limitações devem ser

consideradas. A base de dados utilizada, composta pelos “Cases Campeões” do CESB, representa um conjunto específico de lavouras de alta produtividade, o que pode limitar a capacidade de generalização dos modelos para diferentes realidades produtivas, especialmente em condições menos favoráveis.

Adicionalmente, a avaliação da qualidade do software foi realizada com um número reduzido de especialistas, o que restringe a abrangência das conclusões relacionadas à usabilidade, funcionalidade e desempenho do sistema. A ampliação desse grupo, incluindo profissionais com diferentes níveis de experiência e atuação em distintas regiões, pode contribuir para uma análise mais robusta e representativa.

Outra limitação refere-se aos modelos de aprendizado de máquina utilizados, que, embora amplamente consolidados na literatura, não contemplam abordagens mais recentes baseadas em deep learning e arquiteturas mais avançadas, como redes neurais profundas e modelos baseados em transformers, que vêm sendo explorados em problemas complexos de predição.

Diante disso, como perspectivas para trabalhos futuros, recomenda-se a ampliação da base de dados, a inclusão de novas variáveis agronômicas, a avaliação com um maior número de especialistas e a exploração de técnicas mais avançadas de inteligência artificial. Essas melhorias podem contribuir para aumentar a precisão, a robustez e a capacidade de generalização do sistema.

A expansão do Guiafertil para outras culturas agrícolas, como milho, trigo e algodão, surge também como outra perspectiva para trabalhos futuros sendo um caminho promissor para aumentar sua aplicabilidade. Sugere-se ainda a incorporação de dados provenientes de sensoriamento remoto, como imagens de satélite e drones, o que poderá aprimorar a precisão preditiva dos modelos. Além disso, novas validações com grupos maiores de agrônomos poderão fortalecer a avaliação do software e ampliar sua robustez.

Em síntese, o Guiafertil representa uma contribuição para a agricultura de precisão ao demonstrar que a IA pode ser aplicada de forma prática, acessível e cientificamente embasada na previsão da produtividade agrícola. O software aproxima o produtor e o consultor de uma tomada de decisão mais precisa, baseada em dados reais, e inaugura novas possibilidades para o uso de modelos preditivos no campo. Assim, este trabalho não apenas cumpre seu propósito inicial, mas também abre caminho para avanços futuros no desenvolvimento de tecnologias digitais voltadas ao setor agrícola.

REFERÊNCIAS BIBLIOGRÁFICAS

BUCENE, Luciana C.; RODRIGUES, Luiz H. A. **Utilização de redes neurais artificiais para avaliação de produtividade do solo, visando classificação de terras para irrigação.** Revista Brasileira de Engenharia Agrícola e Ambiental, 2004. https://doi.org/10.1590/S1415-43662004000200025

COMITÊ ESTRATÉGICO SOJA BRASIL (CESB). **Cases campeões**. [s.d.]. Disponível em: <https://www.cesbrasil.org.br/category/cases-campeoes/>. Acesso em: 18 nov. 2024. (CESB, s.d.).

FITZPATRICK, Martin. **Create GUI applications with Python & Qt6**. [S.l.]. Independently published. 1ª edição 2022.

GÉRON, Aurélien. **Mãos à obra: aprendizado de máquina com Scikit-learn, Keras e TensorFlow**. Rio de Janeiro-RJ. Alta Books. 2ª edição, 2021.

GRUS, Joel. **Data Science do zero: noções fundamentais com Python**. Rio de Janeiro-RJ. Alta Books. 2ª edição, 2021.

GUIMARÃES, Edson da Silva. **Aprendizado de máquina aplicado a predição da produtividade da cultura da soja utilizando dados de clima e solo.** Dissertação (Mestrado), Universidade de São Paulo (USP), 2019. https://doi.org/10.11606/D.55.2020.tde-09062020-123106

HALIM, Rico; BUNYAMIN, Hendra. **Analisis penjualan di Cabang Toko Serba Ada dengan algoritma machine Learning**, Strategi, 2023. Disponível em: < http://strategi.itmaranatha.org/ index.php/strategi/article/view/459>. Acesso em: 02/12/2024.

JOSHI, Deepak R.; CLAY, Sharon A. SHARMA,PrakritI; REKABDARKOLAEE, Hossein Moradi; KHAREL, Tulsi; RIZZO, Donna M.; THAPA, Resham; CLAY, David E. **Artificial intelligence and satellite-based remote sensing can be used to predict soybean (*Glycine max*) yield.** Revista International Journal of Agronomy, 2023. https://doi.org/10.1002/agj2.21473

KHALILA, Z.H.; ABDULLAEVA, S.M.. **Neural network for grain yield predicting based multispectral satellite imagery: comparative study.** Revista Elsevier, 2021. https://doi.org/10.1016/j.procs.2021.04.146

LEAL, Aguinaldo José Freitas; MIGUEL, Eder Pereira; BAIO, Fabio H. R.; NEVES, Danilo de Carvalho; LEAL, Ulcilea A. S. **Redes neurais artificiais na predição da produtividade de milho e definição de sítios de manejo diferenciado por meio do solo**. Revista Bragantia, 2015. https://doi.org/10.1590/1678-4499.0140

MUELLER, John; MASSARON, Lucas. **Aprendizado profundo para leigos**. Rio de Janeiro-RJ. Alta Books. 1ª edição, 2020.

PATIL, Sarita; SHRIRAO, Nitin. **Linear regression: A practical implementation in Python**, IJIRCCE 2023. Disponível em: < chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://siddhantica.in/naac/criteria/Criteria_3/3.3.1/Resarch%20Paper/Linear%20regresion%20publish%20paper.pdf >. Acesso em: 02/12/2024.

PEJAK, Branislav; LUGONJA, Predrag; ANTI'C, Aleksandar; PANI'C, Marko; PANDZI'C, Miloš; ALEXAKIS, Emmanouil; MAVREPIS, Philip; ZHOU, Naweiluo; MARKO, Oskar; CRNOJEVI'C, Vladimir. **Soya yield prediction on a within-field scale using machine learning models trained on Sentinel-2 and Soil Data.** Revista MDPI, 2022. https://doi.org/10.3390/rs14092256

REATTO, Adriano; CARVALHO, Arminda M.; SANZONOWICZ, Cláudio; SOUZA, Djalma M. G.; LOBATO, Edson; GALRÃO, Enéas Z.; MENDES, Lêda de C.; CORREIA, João R. C.; SILVA, José E.; ANDRADE, Leide R. M.; VILELA, Lourival; MACEDO, Manuel C. M.; HUNGRIA, Mariangela; LOBO-BURLE, Marília; VARGAS, Milton A. T.; OLIVEIRA, Sebastião A. O.; SPERA, Silvio T.; REIN, Thomaz A.; SOARES, Wilson Vieira. **Cerrado correção do solo e adubação**. Brasília-DF. Embrapa Cerrados. 2ª edição, 2004.

RIBEIRO, Antônio Carlos; GUIMARÃES, Paulo Tacito Gontijo; ALVAREZ, Victor Hugo. **Recomendações para o uso de corretivos e fertilizantes em Minas Gerais: 5ª Aproximação**. Viçosa-MG: Comissão de Fertilidade do Solo do Estado de Minas Gerais - UFV, 1ª edição, 1999.

SANTOS, Letícia Bernabé; GENTRY, Donna; TRYFOROS, Alex; FULTZ, Lisa; BEASLEY, Jeffrey; GENTIMIS, Thanos. **Soybean yield prediction using machine learning algorithms under a cover crop management system**. Revista Elsevier, 2024. https://doi.org/10.1016/j.atech.2024.100442

RIBEIRO, Maxwell M.; GUIMARÃES,Samuel S. **Redes neurais utilizando Tensorflow e Keras**. RE3C - Revista Eletrônica Científica de Ciência da Computação, 2018. Disponível em: <https://revistas.unifenas.br/index.php/RE3C/article/view/231>. Acesso em: 01/10/2024.

SALAM, Mohammed Khalaf. **Optimization of regression models using machine learning: A comprehensive study with scikit-learn**. International Uni-Scientific Research Journal, 2024. https://doi.org/10.59271/s45500.024.0624.16

SHAMMI, Sadia Alam; HUANG, Yanbo; FENG, Gary; TEWOLDE, Haile; ZHANG, Xin; JENKINS, Johnie; SHANKLE, Mark. **Application of UAV Multispectral Imaging to Monitor Soybean Growth with Yield Prediction through Machine Learning**. Revista MDPI, 2024. https://doi.org/10.3390/agronomy14040672

SOARES, Fatima Cibele; ROBAINAM, Adroaldo Dias; PEITERU, Marcia Xavier; RUSSIM, Jumar Luiz. **Predição da produtividade da cultura do milho utilizando rede neural artificial**. Revista Ciência Rural, 2015. https://doi.org/10.1590/0103-8478cr20141524

SOUZA, Vitor Amadeu. **Introdução ao SQLite volume único**. São Paulo-SP. Clube dos Autores. 1ª edição, 2024.

VRIESMANN, Leila Maria; GUIMARÃES, Alaine Margarete; CANTERI, Marcelo Giovanetti; MOLIN, Jose Paulo; CATANEO, Angelo; KOVALECHYN, Danilo. **Analise de resultados obtidos por técnicas de inteligência artificial na mineração de dados de produtividade do solo.** Revista Brasileira de Agrocomputação, 2004. Disponível em:< https://agrocomputacao.deinfo.uepg.br/junho_2004/Arquivos/RBAC.pdf >. Acesso em: 03/05/2023.

YERRABOLU, Venkata Lakshmi; KASIREDDY, Idamakanti Kasireddy; JASMINE, K M; VAMSI, T B Murali Krishna; N, Joshua; V. Shyam Kumar; RAO, DSNM. **Performance comparison of random forest regressor and support vector regression for solar energy prediction**, International Conference on Smart and Sustainable Energy Systems, 2024. https://doi.org/10.1088/1755-1315/1375/1/012013

ZUMSTEIN, Felix. **Python para Excel: um ambiente moderno para automação e análise de dados**. Rio de Janeiro-RJ. Alta Books. 1ª edição, 2024.