ANÁLISE COMPARATIVA DE ALGORITMOS PARA RECOMENDAÇÃO DE PLANTIO NA AGRICULTURA DE PRECISÃO

COMPARATIVE ANALYSIS OF ALGORITHMS FOR PLANTING RECOMMENDATION IN PRECISION AGRICULTURE

Victor Hugo Macedo Silva[1]
Robson de Lacerda Zambroti[2]

RESUMO

Este trabalho realiza uma análise comparativa dos principais algoritmos de *machine learning* utilizados em sistemas de recomendação de plantio. A partir de uma revisão sistemática identificaram-se os algoritmos *Random Forest* (RF), *Support Vectors Machine* (SVM) e *Multilayer Perceptron* (MLP) com utilização de métricas como acurácia, precisão, *recall*, *F1-score*, *Matthews correlation coefficient* (MCC) e *Receiver Operating Characteristics* (ROC). Utilizando dados de características químicas do solo, foi implementada a comparação entre esses algoritmos considerando cenários com normalização, sem normalização e com otimização de hiperparâmetros. O *Random Forest* apresentou os melhores resultados, evidenciando sua robustez e eficácia para sistemas de recomendação de plantio.

Palavras-chave: aprendizado de máquina; floresta aleatória; SVM; MLP; agricultura de precisão.

ABSTRACT

This study conducts a comparative analysis of the main machine learning algorithms used in crop recommendation systems. Through a systematic review, the algorithms identified were Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), utilizing metrics such as accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), and Receiver Operating Characteristics (ROC). Using data on soil chemical properties, these algorithms were compared across scenarios including normalization, no normalization, and hyperparameter optimization. Random Forest delivered the best results, highlighting its robustness and effectiveness for crop recommendation systems.

Keywords: machine learning; random forest; SVM; MLP; precision agriculture.

[1] Discente do curso de Ciência da Computação do Centro Universitário Filadélfia de Londrina - UniFil
[2] Docente do Curso de Ciência da Computação do Centro Universitário Filadélfia de Londrina – UniFil

1 INTRODUÇÃO

Agronegócio é um dos principais pilares da economia brasileira representando R$ 2,45 trilhões no PIB do setor no primeiro trimestre de 2024, com R$ 1,65 trilhões provenientes da agricultura e R$801 bilhões da pecuária (Cepea, 2024). [3]Globalmente movimenta US$2,4 trilhões, sendo essencial para países emergentes e para a segurança alimentar mundial (Katarya *et al*., 2020). No entanto, a agricultura enfrenta desafios como secas, enchentes e mudanças climáticas, que ameaçam a produção exigindo estratégias tecnológicas para mitigar esses impactos.

O crescimento populacional agrava a segurança alimentar, Organização das Nações Unidas (2021)[4] projeta que a população mundial chegará a 9,7 bilhões em 2050, podendo chegar a 10,4 bilhões até 2100. Esse aumento impulsionado pela maior longevidade, urbanização e fluxos migratórios, exige a duplicação da produção de alimentos para garantir a equidade política e social no futuro. No entanto, esse desafio deve ser enfrentado de forma sustentável, pois a expansão agrícola pode causar a destruição de ecossistemas e a degradação de habitats, criando conflitos entre a produção de alimentos e a conservação ambiental (Tilman *et al*., 2002).

A agricultura, além de causar a perda de ecossistemas naturais que fornecem alimentos, fibras, combustíveis e matéria-prima, também impacta negativamente o meio ambiente. O uso excessivo e inadequado de insumos agrícolas pode elevar os níveis de toxinas nas águas subterrâneas e superficiais, contaminando ecossistemas por meio de processos como lixiviação, volatilização e resíduos provenientes de animais e humanos (Tilman *et al*., 2002).

Nesse contexto, a agricultura de precisão surge como uma solução para otimizar o uso de recursos reduzindo desperdícios e impactos ambientais. Baseada na variabilidade espacial e climática, essa técnica permite o manejo mais eficiente de insumos (Bassoi *et al*., 2019). Tecnologias como inteligência artificial (IA) e Internet das Coisas (IoT) têm sido cada vez mais adotadas no setor agrícola para aumentar a

[3] Sumário executivo – pib do agronegócio cepea/cna. Disponível em: <https://www.cepea.esalq.usp.br/upload/kceditor/files/01sut.pib_mar_2024.jul2024-SUMARIO-EXECUTIVO.pdf>. Acesso em: 15 de setembro de 2024.

[4] *Population*. Disponível em: <https://www.un.org/en/global-issues/population> Acesso em: 15 de setembro de 2024.

eficiência e reduzir custos (Katarya *et al*., 2020).

A utilização da inteligência artificial no meio agrícola de modo a duplicar a produção é muito válido, uma vez que ela é capaz de lidar e processar uma enorme quantidade de dados sem desgaste se comparado a nós humanos, ao entender a relação dos dados, é possível promover uma produção mais eficiente e informatizada, levando a um melhor gerenciamento e utilização de água, insumos agrícolas, previsões de safras e plantios, promovendo dessa forma uma prática agrícola mais sustentável e eficiente sem a necessidade de ampliar as áreas agrícolas ou causar muitos gastos monetários.

Desta forma, este trabalho se concentra em uma das subáreas da agricultura de precisão que são os sistemas de recomendação de plantio, que utiliza dados do solo, clima, e fatores ambientais para recomendar a melhor cultura para determinado tipo de solo. O objetivo é comparar os algoritmos de *machine learning* (aprendizado de máquina) mais utilizados identificados na revisão sistemática realizada disponível no PDF[5], de modo a identificar o mais eficaz para a tarefa de recomendação de plantio, oferecendo uma análise e abordagem comparativa que ajude na adoção de inteligencia artificial (IA) no setor agrícola, apoiando decisões mais informadas e sustentáveis.

2 REVISÃO DA LITERATURA

A agricultura tem um papel fundamental no fornecimento de alimentos para a população, a necessidade de dobralá está cada vez aumentando devido ao grande crescimento populacional se tornando um dos maiores desafios enfrentados atualmente (Tilman *et al*., 2002). Em seu estudo, Tilman aponta a importância de alcançar o equilíbrio entre o aumento da produtividade agrícola e a preservação ambiental, especialmente no manejo de nutrientes, uso eficiente de água e na produção de pesticidas. Para lidar com os impactos negativos da agricultura intensivas, práticas como a agricultura de precisão (AP) e agricultura digital (AD) são

[5] Revisão Sistemática Disponível em: https://drive.google.com/file/d/19ruLFjOPNFOZzFc_NqCzPLyamUSfGMAL/view?usp=sharing. Acesso em: 8 ago. 2024.

abordadas no trabalho de Bassoi *et al*. (2019), segundo estes autores, essas abordagens que envolvem automação e digitalização são essenciais para enfrentar desafios como a crescente demanda por alimentos e a escassez de mão de obra especializada, assim como a adoção de práticas agrícolas mais sustentáveis.

Para complementar essa visão, Katarya *et al*. (2020) exploram o uso de técnicas de aprendizado de máquina como *Random Forest* e *Naive Bayes* no contexto de agricultura de precisão. Essas técnicas aumentam a eficiência agrícola ao otimizar o uso de insumos, contribuindo diretamente para a sustentabilidade. Esse ponto é particularmente relevante para países como a Índia que enfrenta dificuldades no setor agrícola, onde o *machine learning* permite decisões mais informadas reduzindo o risco a falhas.

Neste contexto, a escolha adequada de métricas para avaliar os algoritmos de *machine learning* é uma etapa crucial do processo. Em Hossin; Sulaiman (2015) são revisadas as métricas utilizadas para otimizar classificadores, destacando as limitações da acurácia e sugerindo métricas mais robustas como o *F1-score* e o *Receiver Operating Characteristic* (ROC). Em Japkowicz (2013) são analisadas métricas para ambientes com desbalanceamentos de dados, destacando também as limitações da acurácia nesses casos e sugerindo outras mais robustas, enquanto Chicco; Jurman (2020) defendem o *Matthews correlation coefficient* (MCC) como um métrica mais informativa em classificações binárias, demonstrando suas vantagens nesses cenários em comparação a outras. Junior *et al*. (2022), apresentam as principais métricas utilizadas na validação dos algoritmos, ressaltando como as métricas podem refletir o desempenho dos modelos em diferentes contextos.

Ainda para lidar com problemas multiclasses, Grandini; Bagli e Visani (2020) fornecem uma visão geral das principais métricas utilizadas nesse contexto, explicando as generalizações de métricas binárias por meio de médias *macro* e *micro*, essas que na tese de zhang (2021) foi aplicada à métrica *Receiver Operating Characteristic* (ROC), onde permitiu uma avaliação geral do algoritmo por meio da média aritmética de todos os valores de *Area Under The ROC Curve* (AUC). Em (SPEISER *et al*., 2019) comparam em seu trabalho diferentes métodos de seleção de variáveis usando o *Random Forest* (RF), enquanto Fletcher (2009) oferece uma introdução do funcionamento do *Support Vector Machines* (SVM), facilitando a

compreensão de seus fundamentos matemáticos. Costa *et al*. (2023), discutem fundamentos da inteligência computacional, abordando redes neurais, algoritmos evolutivos e sistema *fuzzy*, incluímos estes artigos para entender introdutoriamente o funcionamentos de algoritmos como *Random Forest* (RF), *Support Vector Machines* (SVM) e *Multilayer Perceptron* (MLP).

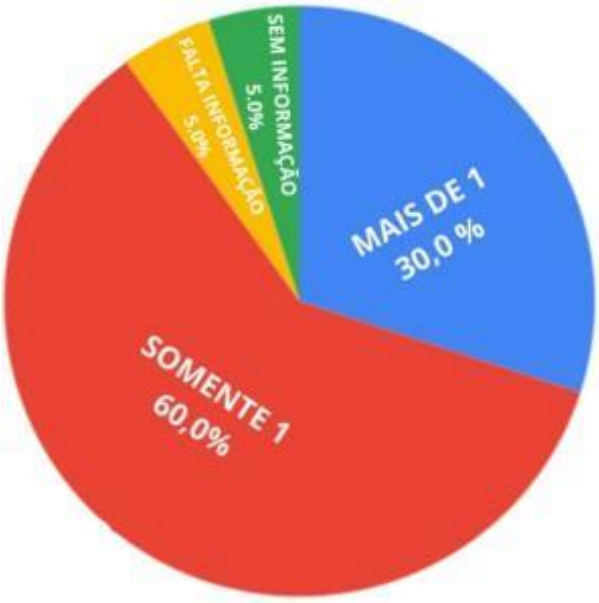
A seleção de *features* é essencial na etapa de pré processamento de dados, em Xia e Yang (2023) e Mariammal *et al*. (2021) é utilizado o *Recursive Feature Elimination* (RFE) para a seleção de *features*, mostrando sua versatilidade com aplicação desde a previsão de culturas até problemas de análise molecular. Prusty, Patnaik e Dash (2022) recomendam a validação cruzada estratificada *(Stratified K-fold)* como alternativa mais robusta ao *K-fold* tradicional, especialmente por garantir proporções balanceadas das amostras em cada divisão dos dados, aplicando com sucesso na previsão do risco de câncer cervical.

Por fim em Beaman; Isah, (2022); Hakim *et al*. (2024) e Hsu (2020) é demonstrado o uso eficaz da técnica *GridSearchCV* para otimização em contextos variado como detecção de *malware* e classificação de textos onde melhorou o desempenho dos algoritmos por meio da otimização de hiperparâmetros.

3 METODOLOGIA

Esta seção detalha os métodos usados neste estudo, incluindo a coleta e análise dos dados. Foi realizada uma revisão bibliográfica de 20 artigos sobre sistemas de recomendação de plantio, utilizando a base de dados *Google Scholar*, que inclui publicações de fontes como *SciELO*, *IEEE Xplore* e *ScienceDirect*. As buscas limitadas aos últimos cinco anos incluíram termos como *"Crop Recommendation System", "Crop Prediction System"e "Crop Prediction Using Machine Learning"*. O objetivo foi identificar as métricas de avaliação e os algoritmos de *machine learning* mais utilizados Figura 1

Figura 1 – Gráfico revisão literatura



Fonte: Os autores.

A revisão revelou que 60% dos estudos analisados baseiam a escolha dos algoritmos de *machine learning* principalmente na métrica de acurácia, confirmando o que (Hossin; Sulaiman, 2015) afirmam sobre sua popularidade na avaliação da generalização dos algoritmos. No entanto, eles alertam que essa métrica pode levar a soluções subótimas, especialmente em bases de dados desbalanceadas, e por possuírem baixa capacidade de discriminar a melhor solução para um classificador. A revisão também destacou que os algoritmos mais utilizados em sistemas de recomendação de plantio foram os baseados em árvores de decisão, como *Random Forest* (RF) e *XGBoost*, que alcançaram altas taxas de acurácia além de redes neurais como o *Multilayer Perceptron* (MLP), que também tiveram bons resultados Quadro 1.

Quadro 1 – Relação entre artigos e algoritmos escolhidos

Artigo	Algoritmo	Conclusão
Crop prediction using machine learning	KNN, Decision Tree, Random Forest	Floresta aleatória, acurácia: 99,32%
Smart Crop Prediction using IoT and MachineLearning	KNN, Decision Tree, SVM	Árvore de decisão acurácia 91%
An Effective Crop Prediction Using Random Forest Algorithm	Random Forest Naive Bayes	Floresta aleatória, acurácia: 95%
Crop Recommender System Using Machine Learning Approach	ANN, SVM, MLR, Random Forest, KNN	Floresta aleatória, acurácia: 95%
Comparative Analysis of Machine Learning Algorithms in The Study of Crop and Crop yield Prediction	SVM, Decision Tree, KNN, Random Forest, Elastic Net, Linear Regression	Decision Tree: 99.87% acurácia
Crop Prediction Model Using Machine Learning Algorithms	Bayes Net, Random Forest, Naïve Bayes, AdaBoost etc...	acurácia de 97,05% com Bayes Net e 97,32% com Random Forest
Improvement of Crop Production Using Recommender System by Weather Forecasts	SVM, ANN, C 5.0, NWP	ANN acurácia 93% Error rate:0.95
Intelligent Crop Recommendation System using Machine Learnin	Decision Tree, KNN,Linear Regression, Naive Bayes, SVM	Rede neural Acurácia 89,88%
IoT Framework for Measurement and Precision Agriculture: Predicting the Crop Using Machine Learning Algorithms	MLP, Decision Table, JRip	MLP mais preciso acurácia: 98,23% ROC: 0.997
Crop Recommendation System using Machine Learning	Decision Tree, Naive Bayes, SVM, Logistic Regression, Random Forest, XGBoost	XGBoost acurácia: 99,31%
Crop Prediction Using Artificial Neural Network and Support Vector Machine	ANN e SVM	ANN Acurácia:86,80%

Fonte: Os autores.

No quadro acima, estão apresentados os principais trabalhos sobre sistemas de recomendação de plantio selecionados na revisão bibliográfica, os algoritmos comparados e as conclusões sobre os modelos mais precisos, conforme a validação dos autores.

4 REVISÃO SISTEMÁTICA

Após a apresentação do pré-projeto deste trabalho, foi sugerido por um dos professores que a realização de uma revisão sistemática seria mais adequada do que uma revisão bibliográfica dos trabalhos dos últimos cinco anos. Deste modo, para a revisão, foi adotado o seguinte Paciente, Intervenção, Comparação, *Outcomes* (resultados) e contexto, também conhecido como PICOC:

- População: agricultores, produtores rurais, fazendeiros e agrônomos.
- Intervenção: Algoritmos de *machine learning*, redes neurais, máquinas de

vetores de suporte (SVM), florestas aleatórias (*Random Forest*) e regressão logística.

- Comparação: Comparação de diferentes algoritmos.
- Resultados: Identificação do algoritmo mais preciso, aumento da produtividade, eficiência no plantio, otimização do uso de recursos e melhoria na tomada de decisão.
- Contexto: Agricultura de precisão, tecnologia agrícola, ambientes agrícolas e sistemas de informação agrícola.

O objetivo desta revisão sistemática foi responder a questões como: quais algoritmos de *machine learning* são mais utilizados para sistemas de recomendação de plantio? Quais métricas são empregadas para avaliar a generalização desses algoritmos? Houve algum método específico de tratativa de dados? Essas perguntas orientaram a seleção das tecnologias a serem incluídas na nossa pesquisa.

5 BUSCA DOS ARTIGOS

A revisão foi conduzida em três bases de dados amplamente reconhecidas no meio acadêmico: *IEEE Xplore*, MDPI e *ScienceDirect*. Foram utilizadas palavras chave gerais, como *"Agriculture". "Farming", "Crop Prediction", "Crop Recommender", "Crop Selection","Crop Suggestion", "Cultivation Suggestion","Planting Recommendaion", "Precision Agriculture", "Precision Farming", "Smart Farming", "Machine Learning", "Automated Learning", "Computational Learning", "ML", "Machine Intelligence", "Predictive Analytics"*. As palavras-chave foram adaptadas para o mecanismo de busca de cada base, garantindo uma cobertura abrangente dos estudos relevantes.

6 CRITÉRIO DE EXCLUSÃO

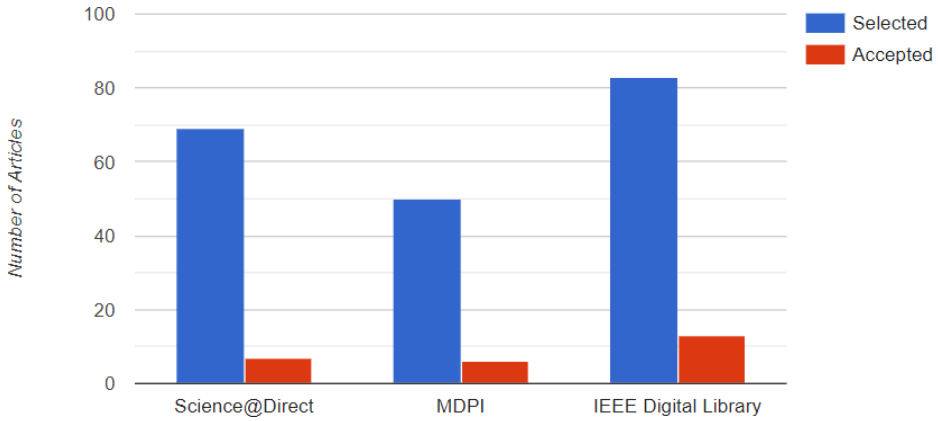
Os artigos foram selecionados ou excluídos com base nos seguintes critérios: (1) artigos não publicados nos últimos cinco anos, (2) falta de acesso completo ao trabalho, (3) artigos fora do escopo da pesquisa, (4) artigos não publicados em inglês,

(5) artigos duplicados e (6) trabalhos que não tratam de sistemas de recomendação de plantio ou que não utilizam dados relacionados a informações e características do solo.

7 RESULTADOS

Os resultados da revisão sistemática podem ser visualizados no gráfico da Figura 2.

Figura 2 – Gráfico revisão sistemática



Fonte: Os autores

Para o *ScienceDirect*, 69 artigos foram inicialmente selecionados dos quais 7 foram aceitos; no MDPI, 50 artigos foram identificados com 6 aceitos; e no *IEEE Xplore*, 83 artigos foram selecionados resultando em 13 aceitos. No total, 26 artigos foram incluídos na revisão sobre sistemas de recomendação de plantio.

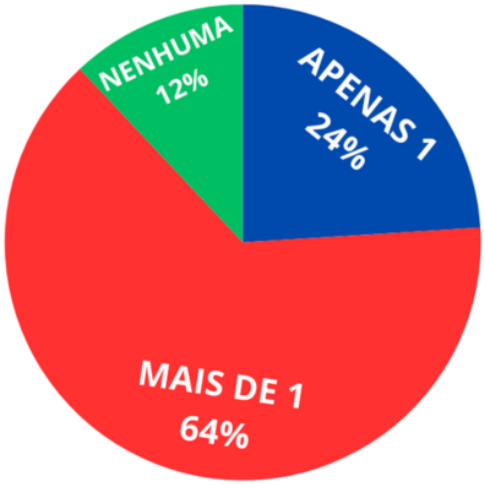
Conforme os dados apresentados, os algoritmos mais utilizados nos estudos analisados foram o *Support Vector Machine* (SVM), os baseados em árvores de decisão, como o *Random Forest* (RF), e as redes neurais, com destaque para o *Multlayer Perceptron* (MLP). Sua escolha se justifica pela comprovada eficácia na resolução de problemas de classificação complexos, como indicado na revisão sistemática disponível no PDF[6] e na revisão de literatura (Quadro 1). O *Random*

[6] Revisão Sistemática. Disponível em: https://drive.google.com/file/d/19ruLFjOPNFOZzFc_NqCzPLyamUSfGMAL/view?usp=sharing. Acesso em: 8 ago. 2024.

Forest (RF) se destaca por sua robustez e capacidade de lidar com dados altamente dimensionados, o *Support Vector Machine* (SVM) por sua habilidade em trabalhar com dados de margens bem definidas e flexibilidade diante de dados complexos, e o *Multlayer Perceptron* (MLP) por sua eficiência na generalização e captura de padrões complexos. Além disso, observou-se uma mudança significativa nas métricas utilizadas para validar a precisão dos algoritmos em comparação à revisão bibliográfica anterior Figura 1.

A revisão sistemática mostrou que 64% dos autores validam algoritmos de *machine learning* com mais de uma métrica de avaliação, o que (CHICCO; JURMAN, 2020) em seu estudo de caso mostrando a robustez da métrica *Matthews correlation coefficient* (MCC) mostra que se basear somente em uma métrica pode resultar em uma análise otimista e imprecisa do desempenho do algoritmo. Já 24% dos estudos se baseiam apenas na acurácia, que segundo (JAPKOWICZ, 2013), tende a favorecer a classe majoritária, ignorando a minoritária e resultando em avaliações imprecisas. Além disso, 12% dos trabalhos não apresentaram dados que validassem suas abordagens de *machine learning* Figura 3.

Figura 3 – Gráfico resultados revisão sistemática



Fonte: Os autores

8 PROCEDIMENTOS EXPERIMENTAIS

Após definir as tecnologias usadas na pesquisa, os algoritmos foram executados em um conjunto de dados retirado do *Kaggle*, estes dados são referentes a informações numéricas químicas do solo como o nitrogênio, fósforo, potássio, PH (acidez), temperatura, umidade, precipitação, água disponível por ano, e o tipo de cultura recomendada segundo essas características, sendo representado por 9 colunas para cada uma dessas variáveis chegando a um total de 2222 linhas de registros. Eles foram limpos tratando todos os valores nulos ou vazios. As variáveis numéricas foram normalizadas para que todas estivessem na mesma escala, uma condição necessária para algoritmos como *Support Vector Machine* (SVM) ou *Multilayer Perceptron* (MLP) logo que são sensíveis à escalabilidade de dados como observado nos testes. O método *Recursive Feature Elimination* (RFE) também foi empregado para retirar as *features* menos importantes da base de dados, diminuindo a complexidade do problema.

Os algoritmos foram submetidos em 4 cenários: (1) dados não normalizados e hiperparâmetros padrão, aqui buscamos observar como a ausência de pré processamento pode afetar o desempenho dos modelos, principalmente em relação a escalabilidade dos dados, (2) não normalizados com hiperparâmetros otimizados, avaliamos se com a otimização é possível mitigar o impacto da não normalização, (3) dados normalizados e hiperparâmetros padrão, exploramos a importância da normalização dos dados, principalmente para os algoritmos SVM e MLP, e por fim, (4) base de dados normalizada com hiperparâmetros otimizados, analisamos o impacto dessa combinação na previsão dos algoritmos.

As métricas de avaliação utilizadas seguiram as recomendações da literatura garantindo uma análise robusta e comparativa. Para otimização dos hiperparâmetros, utilizamos o *GridSearchCV*, permitindo uma busca eficiente pelas melhores combinações. As tecnologias aplicadas neste trabalho serão detalhadas na próxima seção.

9 TECNOLOGIA

As escolhas dos algoritmos neste trabalho foram baseadas na revisão sistemática que identificou os mais utilizados em sistemas de recomendação de plantio na agricultura de precisão. Dos 26 artigos selecionados, 5 utilizaram árvores de decisão principalmente *Random Forest* (RF) (Katarya *et al*., 2020; Savla *et al*., 2020; Rao *et al*., 2022; P; R, 2021; Islam *et al*., 2023) , 4 aplicaram o algoritmo *Support Vector Machine* (SVM) (Sizan *et al*., 2023; M.; Megalingam, 2019; Nagaraja *et al*., 2019; Dash; Dash; Biswal, 2021) e 4 usaram redes neurais, com destaque para o *Multilayer Perceptron* (MLP) (A *et al*., 2021; P; R, 2021; Ranjan; Garg; Rai, 2022; Bakthavatchalam *et al*., 2022), o anexo com todos os dados da revisão sistemática pode ser acessado pelo arquivo[7]. A seguir, será apresentada uma breve descrição de cada um desses algoritmos

- ***Random Forest* (RF):** Speiser *et al*. (2019) afirmam que os algoritmos de floresta aleatória consistem em múltiplas árvores de decisão que realizam divisões binárias nas variáveis preditoras. As saídas dessas árvores são agregadas, resultando em previsões mais precisas do que um modelo de árvore único. O principal benefício do RF é sua capacidade de lidar com muitas variáveis e sua facilidade de uso.

- ***Support Vector Machine* (SVM):** O *Support Vector Machine* (SVM) é um algoritmo de classificação e regressão que busca um hiperplano que separa as classes maximizando a margem entre os vetores de suporte, melhorando a generalização do modelo. Ideal para dados lineares, o SVM pode usar *kernels* como linear, polinomial e *Radial Basis Function*, para lidar com casos não lineares (Fletcher, 2009).

- ***Multilayer Perceptron* (MLP):** Redes neurais inspiradas no sistema nervoso, são compostas por neurônios artificiais que resolvem problemas matemáticos. O *Multi Layer Perceptrons* (MLP), uma das arquiteturas mais conhecidas, organiza esses neurônios em camadas encadeadas *feedforward*, onde cada

[7] Revisão Sistemática. Disponível em: https://drive.google.com/file/d/19ruLFjOPNFOZzFc_NqCzPLyamUSfGMAL/view?usp=sharing. Acesso em: 8 ago. 2024.

camada envia sinais para a próxima. A camada de entrada recebe os dados, a camada oculta mapeia o sinal não linear para um outro e a camada de saída gera os resultados (Costa *et al*., 2023).

10 MÉTRICAS DE AVALIAÇÃO

Os autores Grandini; Bagli; Visani (2020) destacam a utilidade de diversas métricas para avaliar classificadores multiclasse, eficazes tanto na comparação de modelos quanto na análise de diferentes configurações. Neste trabalho, utilizamos as métricas mais citadas na revisão sistemática reconhecidas por sua robustez.

11 ACURÁCIA

A acurácia é uma métrica simples e amplamente aplicável a problemas multiclasse e *multilabels* (Hossin; Sulaiman, 2015). Ela representa o percentual de previsões corretas feitas pelo algoritmo em relação ao total de observações. A acurácia pode ser calculada pela fórmula abaixo, considerando os verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN).

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

Fonte: Adaptado de Grandini, Bagli e Visani (2020, p. 3).

12 PRECISÃO *MACRO* E RECALL *MACRO*

A precisão calcula a porcentagem de verdadeiros positivos (TP) corretamente classificados em relação ao total de positivos previstos (TP + FP) (Junior *et al*., 2022). O *recall* mede a capacidade do algoritmo de identificar todas as unidades positivas, sendo a razão entre verdadeiros positivos (TP) e o total de positivos reais (TP+ FN). (Japkowicz, 2013; Grandini; Bagli; Visani, 2020) destacam que essas métricas são

utilizadas em classificação binária. Para problemas multiclasse, os autores sugerem a média *macro*, que calcula a média aritmética da precisão e *recall* para cada classe atribuindo peso igual a todas elas.

$$Precisão_k = \frac{TP_k}{TP_k + FP_k} \qquad Precisão\ macro = \sum_{k=1}^{k} \frac{Precisão_k}{K}$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \qquad Recall\ macro = \sum_{k=1}^{k} \frac{Recall_k}{K}$$

Fonte: Adaptado de Grandini, Bagli e Visani (2020, p. 7).

13 F1-SCORE MACRO

Em Hossin; Sulaiman (2015) é afirmado que o *F1-Score* é um bom discriminador de algoritmos, superando a acurácia na otimização dos classificadores. É definido como a média harmônica entre a precisão e o *recall*, variando de 0 (pior desempenho) a 1 (melhor desempenho). Sua extensão para o *F1-Score macro* é a média harmônica das versões *macro* de precisão e *recall* (Grandini; Bagli; Visani, 2020).

$$Macro\ F1 - score = 2x\ \frac{Precisão\ macro\ x\ Recall\ macro}{precisão\ macro + recall\ macro}$$

Fonte: Adaptado de Grandini, Bagli e Visani (2020, p. 7).

14 *MATTHEWS CORRELATION COEFFICIENTS* (MCC)

A métrica *Matthews correlation coefficients* (MCC) avalia o desempenho do algoritmo considerando todos os elementos da matriz de confusão, variando de -1 (pior) a +1 (melhor), com 0 representando aleatoriedade (Chicco; Jurman, 2020). Em problemas multiclasse, o MCC é calculado com base em uma matriz que abrange todas as classes considerando os elementos corretamente classificados (Grandini; Bagli; Visani, 2020).

$$MCC = \frac{c\ x\ s - \sum_k^k p_{k\,x}\, t_k}{\sqrt{\left(s^2 - \sum_k^k p_k^2\right)\left(s^2 - \sum_k^k t_k^2\right)}}$$

$c = \sum_k^k C_{kk}$: Total de elementos corretamente previsto.
$s = \sum_i C_j\, C_{ij}$: O total de elementos.
$p_k = \sum_i C_{ki}$: Número de vezes que a classe k foi prevista.
$t_k = \sum_i C_{ik}$: Número de vezes que a classe k realmente ocorreu.

Fonte: Adaptado de Grandini, Bagli e Visani (2020, p. 11)

15 *RECEIVER OPERATING CHARACTERISTICS* (CURVA ROC)

A curva ROC avalia a capacidade do algoritmo de distinguir entre classes (Junior *et al*., 2022). Embora seja comumente aplicada em classificação binária, pode ser adaptada para multiclasse gerando k curvas ROC e combinando-as com a média *macro* (Zhang, 2021). Embora que a curva ROC seja intuitiva, temos o valor de *Area Under Curve* (AUC) que varia de 0,5 a 1, sendo 1 o melhor caso. O calculado da curva ROC é representado a seguir:

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + FN}$$

Fonte: Adaptado Mengze Zhang (2021, p. 4)

16 *RECURSIVE FEATURE ELIMINATION* (RFE)

Segundo Xia; Yang, (2023) e Mariammal *et al*. (2021) , o *Recursive Feature Elimination* (RFE) é um método recursivo que elimina características com base na importância atribuída por um estimador, removendo as menos relevantes até que apenas as mais importantes permaneçam. Em nosso trabalho, utilizamos o *Random Forest* (RF) para seleção de *features* devido ao seu sucesso comprovado em diversos estudos como Xia; Yang (2023), Mariammal *et al*. (2021) e Chen *et al*. (2018), esse último destaca as vantagens do *Random Forest* (RF) como um estimador imparcial e eficaz na medição da importância das características.

17 *STRATIFIED K-FOLD*

Segundo Prusty; Patnaik; Dash (2022), a validação *K-fold* estratificada é ideal para problemas de classificação, pois garante que cada *fold* mantenha a proporção das classes resolvendo o problema de distribuições desbalanceadas no *K-fold* tradicional. Nessa abordagem, os dados são divididos em k subconjuntos e o modelo é treinado k vezes, usando um *fold* como teste e os outros k-1 para treinamento.

18 *GRIDSEARCH*

O *GridSearch*, como descrito por (Hakim *et al*., 2024), otimiza os hiperparâmetros de um modelo de *machine learning* ao testar todas as combinações possíveis dentro de uma grade, buscando maximizar o desempenho do classificador. No entanto, como (HSU, 2020) aponta, essa técnica é intensiva e demorada devido à avaliação de todas as combinações. Escolhemos o *GridSearch* pelos bons resultados observados em estudos como (Sizan *et al*., 2023; Islam *et al*., 2023; Beaman; Isah, 2022; Hakim *et al*., 2024; Hsu, 2020).

19 NORMALIZAÇÃO DOS DADOS

Por se tratar de uma base de dados contendo características químicas do solo muitas delas provenientes de sensores IoT, os dados estão em escalas diferentes tornando - se necessário trata-las de modo a deixar todas as variáveis em uma escala comum, principalmente se tratando de algoritmos como SVM e o MLP que são sensíveis.

O processo de normalização utilizado neste trabalho foi um método presente na biblioteca *Scikit-learn* do *python* chamado *StandardScaler*. Esse método ajusta os dados para que tenham média igual a 0 e desvio padrão igual a 1, ou seja, deixa todos os valores em nossa base de dados em uma escala padronizada Tabela 1.

Tabela 1 – Base de dados - Sem aplicação *StandardScaler* vs. Com aplicação *StandardScaler*

Index	N	P	K	Temperatura (°C)	Umidade (%)
Sem *StandardScaler*					
0	90	42	43	20.879744	82.002744
1	85	58	41	21.770462	80.319644
2	60	55	44	23.004459	82.320763
3	74	35	40	26.491096	80.158363
4	78	42	42	20.130175	81.604873
Com *StandardScaler*					
0	0.323433	-0.411340	-0.213914	-0.889405	0.488736
1	0.261639	-0.057588	-0.247171	-0.722967	0.412844
2	-0.047332	-0.123917	-0.197286	-0.492384	0.503076
3	0.125692	-0.566107	-0.263799	0.159124	0.405571
4	0.175127	-0.411340	-0.230542	-1.029468	0.470796

Fonte: Os autores

Na tabela acima conseguimos comparar as 5 primeiras colunas e linhas de nossa base de dados que foi utilizada, com os valores sem e com o processo de normalização do *StandardScaler*.

20 RESULTADOS

A tabela 2 a seguir apresenta o desempenho dos algoritmos nos 10 *Folds* da validação cruzada tanto no cenário não normalizado quanto o normalizado sem otimização.

Tabela 2 – Resultados Comparativos - Não Normalizado e Não Otimizado vs. Normalizado Não Otimizado

Métricas	RF	SVM	MLP
Não Normalizado e Não Otimizado			
Accuracy	0.9995 ± 0.0014	0.2937 ± 0.1982	0.9509 ± 0.0318
F1-Score	0.9995 ± 0.0014	0.2487 ± 0.2119	0.9450 ± 0.0376
Precision	0.9996 ± 0.0012	0.2585 ± 0.2153	0.9498 ± 0.0367
Recall	0.9995 ± 0.0014	0.2873 ± 0.1999	0.9506 ± 0.0319
MCC	0.9995 ± 0.0014	0.3533 ± 0.1773	0.9493 ± 0.0326
AUC	1.0000 ± 0.0000	0.4482 ± 0.0015	0.9887 ± 0.0159
Tempo	3.2254s	10.7271s	7.5260s
Normalizado Não Otimizado			
Accuracy	0.9986 ± 0.0021	0.9748 ± 0.0090	0.9856 ± 0.0069
F1-Score	0.9986 ± 0.0021	0.9748 ± 0.0090	0.9855 ± 0.0070
Precision	0.9988 ± 0.0019	0.9765 ± 0.0085	0.9869 ± 0.0068
Recall	0.9986 ± 0.0021	0.9745 ± 0.0091	0.9855 ± 0.0070
MCC	0.9986 ± 0.0022	0.9737 ± 0.0094	0.9850 ± 0.0072
AUC	1.0000 ± 0.0000	0.9996 ± 0.0002	0.9999 ± 0.0001
Tempo	3.0160s	2.5620s	11.4090s

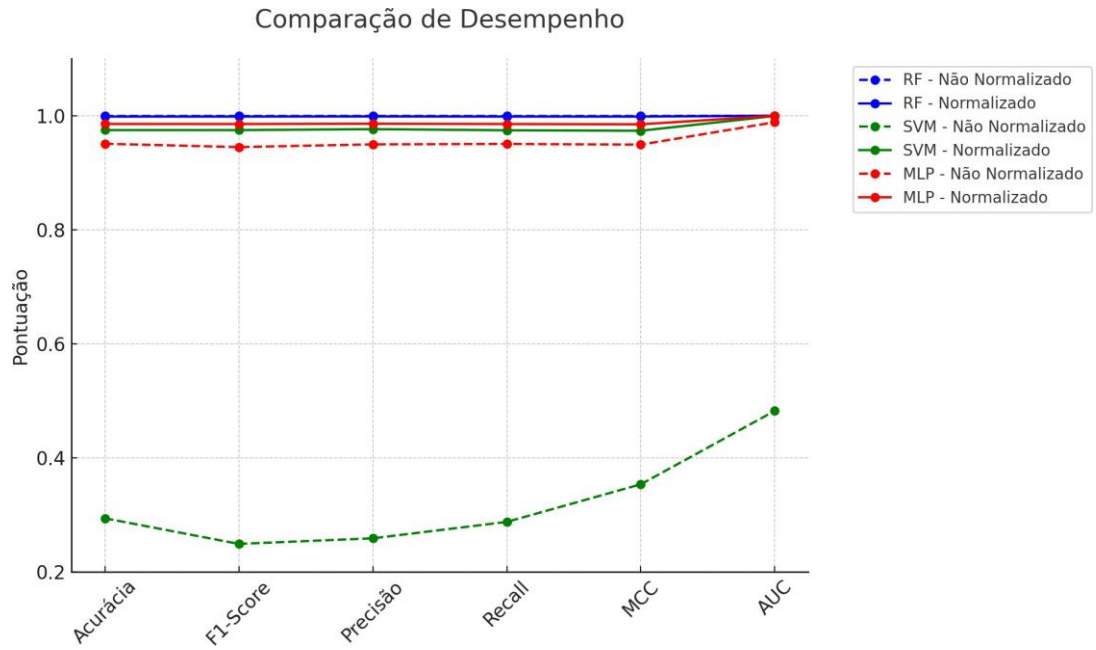
Fonte: Os autores

Analisando os dados na tabela 2 e o gráfico presente na figura 4, observa- se que o RF obteve o melhor desempenho em comparação aos demais algoritmos, tanto no cenário normalizado quanto o não normalizado, isso destaca sua robustez e capacidade de lidar com a escalabilidade dos dados. Em ambos os cenários seu desempenho manteve equilibrado com valores a partir de 0.9995 em todas as métricas com uma queda mínima no cenário com dados normalizados. O gráfico abaixo figura 4 reflete seus altos valores nas métricas, com a linha do cenário normalizado e não normalizado sendo apresentado próximos do valor 1.0 referente a escala máxima.

Em contrapartida, o SVM apresentou o pior desempenho em todas as métricas no cenário não normalizado, isso é evidenciado pela variação de seu desempenho com valores de 0.2487 a 0.4482 Tabela 2, já no cenário otimizado o SVM obteve um impacto preditivo muito positivo. Com base nos valores apresentados seus resultados variaram de 0.9748 (menor valor) à 0.9996 (maior valor) com a sua linha (cenário normalizado) sendo representada próximo do 1.0, um desempenho muito alto em relação a falta de normalização como refletido na Figura 4. O MLP por sua vez demonstrou um desempenho intermediário entre o RF e o SVM, onde a sua linha no

cenário normalizado e não otimizado situou-se entre os resultados desses dois algoritmos.

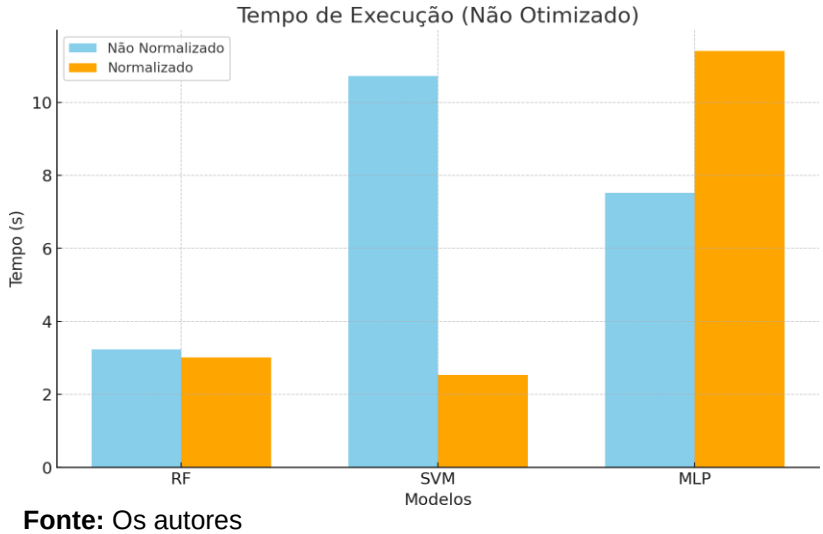
Figura 4 – Gráfico comparação não otimizado



Fonte: Os autores

Quanto ao tempo de treinamento, para o RF obtivemos uma redução de 3.2254s para 3.0160s demonstrando um tempo estável independentemente da normalização dos dados, O SVM obteve uma melhoria de 10.7271s (sem a normalização) para 2.5620s (com normalização) e o MLP devido a erros de convergência Figura 7 obteve um aumento no tempo de treinamento de 7.5286s para 11.4090s. A relação da melhoria do tempo dos algoritmos podem ser observados no gráfico abaixo Figura 5.

Figura 5 – Gráfico comparação de tempo (não otimizado)



Fonte: Os autores

Como indicado na Figura 6, o SVM não conseguiu prever amostras para algumas classes em 9 dos 10 *folds*, resultando em precisão indefinida para as mesmas o que reforça a importância da normalização, o mesmo erro aconteceu para o MLP que falhou em prever corretamente as classes em 2 dos 10 *folds.*

Figura 6 – Erro de classificação

```
C:\Users\Claudio e Victor\PycharmProjects\comparandoPadraoNaoNormalizado\envNovo\Lib\site-packages\sklearn\metrics\
_classification.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples.
Use `zero division` parameter to control this behavior.  warn prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

Fonte: Os autores

Embora o MLP tenha superado o SVM em uma base de dados não normalizada com valores variando de 0.9450 à 0.9887, ainda enfrentou problemas de convergência aos dados para *Maximum Iterations* (200) apresentado na figura 7. Isso significa que o MLP não conseguiu convergir em 200 iterações, indicando que a rede não ajustou corretamente seus pesos para minimizar o erro entre as previsões e os valores reais, esse problema pode ser atribuído à falta de normalização dos dados.

Figura 7 – Erro de convergência

```
C:\Users\Claudio e Victor\PycharmProjects\comparandoPadraoNaoNormalizado\envNovo\Lib\site-packages\sklearn\neural_network\
_multilayer_perceptron.py:690: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization
hasn't converged yet. warnings.warn(
```

Fonte: Os autores

A normalização beneficiou a rede neural MLP que conseguiu prever todas as classes nos *fold's* não apresentando mais os erros anteriores de classificação Figura 6, isso se da pelo fato que a normalização impede que *features* de maior magnitude dominem o processo de aprendizado.

O aumento do tempo de treinamento ocorrido no MLP pode ser atribuído aos problemas de convergência Figura 7 sugerindo que ajustes adicionais nos hiperparâmetros ou número maior de iterações são necessárias, o que será explorado nos testes seguintes.

21 CENÁRIO NORMALIZADO E NÃO NORMALIZADO COM OTIMIZAÇÃO

Neste cenário iremos analisar os algoritmo otimizados tanto na base de dados normalizado como não normalizada, para verificarmos se a otimização pode mitigar os problemas dos algoritmos Tabela 3.

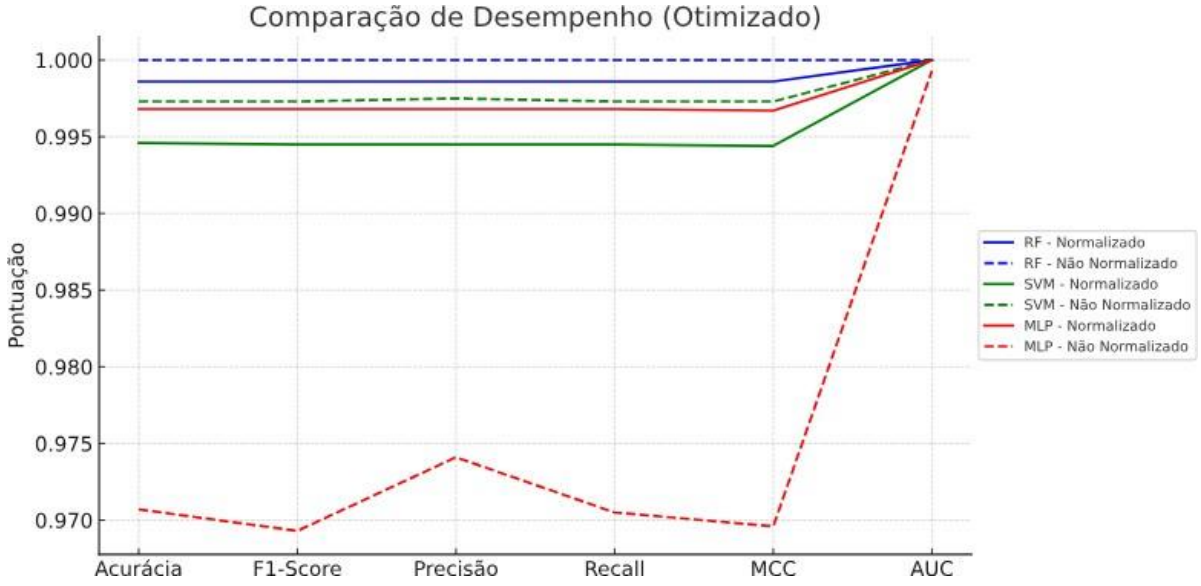
Tabela 3 – Resultados Comparativos - Otimizado Não Normalizado vs. Normalizado e Otimizado

Métricas	RF	SVM	MLP
Otimizado Não Normalizado			
Accuracy	1.0000 ± 0.0000	0.9973 ± 0.0030	0.9707 ± 0.0174
F1-Score	1.0000 ± 0.0000	0.9973 ± 0.0030	0.9693 ± 0.0198
Precision	1.0000 ± 0.0000	0.9975 ± 0.0028	0.9741 ± 0.0139
Recall	1.0000 ± 0.0000	0.9973 ± 0.0030	0.9705 ± 0.0175
MCC	1.0000 ± 0.0000	0.9972 ± 0.0031	0.9696 ± 0.0178
AUC	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9993 ± 0.0005
Tempo	6.9130s	1.1800s	16.2120s
Normalizado e Otimizado			
Accuracy	0.9986 ± 0.0021	0.9946 ± 0.0048	0.9968 ± 0.0041
F1-Score	0.9986 ± 0.0021	0.9945 ± 0.0050	0.9968 ± 0.0041
Precision	0.9988 ± 0.0019	0.9949 ± 0.0046	0.9971 ± 0.0037
Recall	0.9986 ± 0.0021	0.9945 ± 0.0049	0.9968 ± 0.0041
MCC	0.9986 ± 0.0022	0.9944 ± 0.0050	0.9967 ± 0.0042
AUC	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Tempo	3.0350s	0.8990s	6.8759s

Fonte: Os autores

Para o RF utilizamos a grade *"criterion: gini", "max_depth: 10", "max_features: None", "min_samples_leaf: 1", "min_samples_split: 2", "n_estimators: 100"*. Embora a otimização dos hiperparâmetros não tenha causado grandes mudanças, o modelo alcançou uma porcentagem perfeita nas métricas, demonstrando que o mesmo se ajustou completamente aos dados Figura 8.

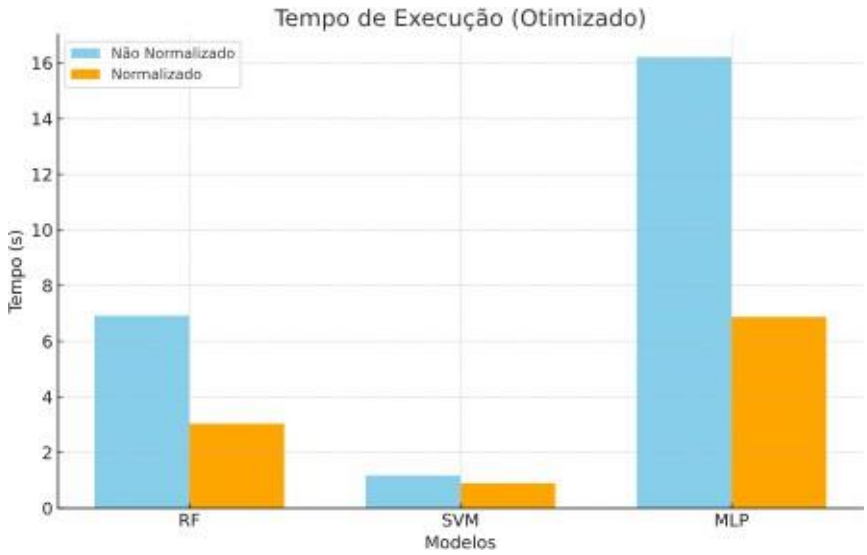
Figura 8 – Comparação não normalizado e normalizado com otimização



Fonte: Os autores

Com o aumento da complexidade, em comparação ao cenário sem otimização e normalização, o tempo aumentou de 3.0160s (Tabela 2) para 6.9130s (Tabela 3), reflexo de um modelo mais complexo com o ajuste fino dos hiperparâmetros. Para os dados normalizados a configuração mudou para *"criterion: gini", "max_depth: 10","max_features: sqrt", "min_samples_leaf: 1", "min_samples_split: 10", "n_estimators: 100"*, houve uma pequena redução nas métricas de 1.000 ± 0.0000 para 0.9986 ± 0.0021 atribuída à maior generalização e estabilidade do RF em dados brutos. Mesmo assim, o algoritmo manteve uma acurácia acima de 99.85% além de que normalização ajudou a reduzir o tempo de treinamento de 6.9130s para 3.0350 como mostrado na Figura 9.

Figura 9 – Comparação tempo não normalizado e normalizado com otimização



Fonte: Os autores

No caso do SVM a otimização foi fundamental. Com a grade *"C: 0.01", "degree: 2", "gamma: 0.01", "kernel: poly"*, a acurácia aumentou de 0.2937 ± 0.1982 (Tabela 2, cenário não normalizado e não otimizado) para para 0.9973 ± 0.0030 (Tabela 3, cenário não normalizado e otimizado), corrigindo os erros de classificação e reduzindo o tempo de treinamento de 10.7271s para 1.1800s. Esses resultados demonstram que a escolha adequada de *kernel* e parâmetros foi essencial para permitir que o modelo convergisse mais rapidamente aos dados e o tornasse mais robusto. No cenário otimizado e normalizado a grade *"C: 10", "degree: 2", "gamma: scale", "kernel: linear"*, apresentou uma leve queda nas métricas Tabela 3 visível pela linha sólida na Figura 8, contudo o modelo ainda apresentou bons resultados com valores a partir de 0.9944. A normalização, junto com a otimização, também reduziu o tempo de treinamento para 0.8990s, conforme Figura 9.

Para o MLP, a acurácia aumentou de 0.9509 ± 0.0318 (Tabela 2, cenário não normalizado e não otimizado) para 0.9707 ± 0.0174 (Tabela 3, cenário não normalizado e otimizado) com a grade *"activation': tanh", "alpha: 0.01", "hidden_layer_sizes: (100,)", "learning_rate_init: 0.0005", "max_iter: 1000", "solver: adam"*. A otimização corrigiu erros e melhorou o desempenho geral, embora o tempo de execução tenha aumentado de 7.5260s para 16.2120s, refletindo a maior complexidade do modelo após o ajuste dos hiperparâmetros. No cenário normalizado

e otimizado, o MLP foi o que mais se beneficiou com a configuração *"activation: tanh", "alpha: 0.01", "hid- den_layer_sizes: (50, 50)", "learning_rate_init: 0.0005", "max_iter: 200", "solver: lbfgs"*. alcançando as maiores porcentagem em relação aos testes anteriores, até mesmo no cenário não normalizado com otimização como observado no gráfico da Figura 8, além de reduzir o tempo de execução de 16.2120s para 6.8759s (Figura 9) alcançando o seu tempo mais rápido para a convergência dos dados, permitindo uma arquitetura mais simples e menos custosa para obter um desempenho satisfatório.

22 CONCLUSÃO

Após a normalização dos dados ajustes nos algoritmos simplificaram sua configuração, o RF foi reconfigurado para *"max_features: sqrt" e "min_samples_split: 10"* limitando as divisões e a quantidade de variáveis utilizadas em cada nó, no SVM o ajuste para *"C = 10", "kernel = linear" e "gamma = scale"* aumentado a penalização por erros e assumindo classes linearmente separáveis, já o MLP com camadas reduzidas e menor número de iterações junto com um *kernel* mais simples, otimizou-se o tempo de treinamento. Essas modificações resultaram em algoritmos mais simples e eficiente com melhor capacidade generativa, como foi demonstrado nos resultados.

Este estudo contribui para a agricultura de precisão ao fornecer uma análise comparativa dos principais algoritmos de *machine learning*, e evidenciar a eficiência da média *macro* na validação dos algoritmos, com isso, os resultados deste trabalho alcançaram o objetivo proposto por essa pesquisa que foi identificar o algoritmo mais preciso para a recomendação de plantio, comprovando que o *Random Forest* (RF) é o algoritmo mais robusto para a tarefa, atingindo 0.99 em todas as métricas nos 4 cenários que foram testados. Esse desempenho, segundo (Grandini; Bagli; Visani, 2020) evidencia sua precisão, com baixo índice de falsos positivos e alta correlação entre previsões e rótulos reais. Além disso, a AUC *macro* perfeita como destacado por (Zhang, 2021) confirma a capacidade do RF em separar eficientemente as classes. A resistência a variações de escala do RF é fundamental, pois operações agrícolas

modernas geram dados de diferentes sensores IoT (Liakos *et al*., 2018) e Integrar esse volume de dados com o conhecimento agronômico é essencial para aumentar a produção agrícola, conforme (Bassoi *et al*., 2019).

O *Support Vector Machine* (SVM) também apresentou um desempenho notável após a normalização com o tempo de treinamento reduzido para 0.8990 segundos, indicando rápida convergência. Em cenários que exigem rapidez no treinamento com entrada de novos dados em tempo real, o SVM otimizado pode ser uma excelente opção.

Pesquisas futuras podem explorar algoritmos avançados de *Deep Learning* para superar desafios de padronização e acesso a dados conforme apontada por (Saraiva *et al*., 2024). O sucesso desses algoritmos depende e exige grandes volumes de dados em bom estado, além de um alto esforço computacional, especialmente ao se utilizar técnicas como o *gridSearchCV* (HSU, 2020). Métodos que combinem a robustez do RF com a velocidade do SVM, além da inclusão de variáveis climáticas e dados em tempo real podem fortalecer esses sistemas.

AGRADECIMENTOS

Agradeço primeiramente a Deus e à Virgem Maria pela força constante nesta jornada. Sou grato à minha família, especialmente à minha mãe Ana Maria, meu pai Claudio Alves e meu irmão Claudio Henrique por todo apoio incondicional além das orações que me ajudaram a lidar com os desafios deste percurso. Deixo um agradecimento especial ao meu avô Benedito Zeferino de Macedo que vim a perder durante a realização desse trabalho, cujas palavras me inspiraram e me deram força.

Expresso minha gratidão ao meu primo Luis Fernando por me ajudar a realizar o sonho de me formar, e a meus amigos André Oliveira, Luana Martins, Laura Arantes, Ana Carolina, Victor Hugo Rodrigues, Silvio Lucas, Mateus Martins e Gabriel Ferreira, que foram apoio essencial durante o curso. Agradeço também ao meu orientador, Prof. Robson de Lacerda Zambroti, por sua orientação e paciência. Agradeço também a todos que leram e contribuíram de alguma forma para este trabalho.

REFERÊNCIAS

A, P. *et al*. Intelligent crop recommendation system using machine learning. In: *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. [S.l.]: IEEE, 2021.

BAKTHAVATCHALAM, K. *et al*. Iot framework for measurement and precision agriculture: Predicting the crop using machine learning algorithms. *Technologies*, MDPI AG, v. 10, n. 1, p. 13, jan. 2022. ISSN 2227-7080.

BASSOI, L. H. *et al*. Agricultura de precisão e agricultura digital. *TECCOGS: Revista Digital de Tecnologias Cognitivas*, Pontifical Catholic University of Sao Paulo (PUC-SP), n. 20, maio 2019. ISSN 1984-3585.

BEAMAN, C.; ISAH, H. *Anomaly Detection in Emails using Machine Learning and Header Information*. [S.l.]: arXiv, 2022.

CHEN, Q. *et al*. Decision variants for the automatic determination of optimal feature subset in rf-rfe. *Genes*, MDPI AG, v. 9, n. 6, p. 301, jun. 2018. ISSN 2073-4425.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, Springer Science and Business Media LLC, v. 21, n. 1, jan. 2020. ISSN 1471-2164.

COSTA, L. *et al*. Multilayer perceptron. *Introduction to Computational Intelligence*, v. 105, 2023. Disponível em: https://purehost.bath.ac.uk/ws/files/279724412/Introduction_to_Computational_Intelligence.pdf#page=134.

DASH, R.; DASH, D. K.; BISWAL, G. Classification of crop based on macronutrients and weather data using machine learning techniques. *Results in Engineering*, ElsevierBV, v. 9, p. 100203, mar. 2021. ISSN 2590-1230.

FLETCHER, T. Support vector machines explained. *Tutorial paper*, v. 1118, p. 1–19, 2009. Disponível em: https://www.csd.uwo.ca/~xling/cs860/papers/SVM_Explained.pdf.

GRANDINI, M.; BAGLI, E.; VISANI, G. *Metrics for Multi-Class Classification: an Overview*. [S.l.]: arXiv, 2020.

HAKIM, L. *et al*. Optimzing android program malware classification using gridsearchcv optimized random forest. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Universitas Muhammadiyah Malang, maio 2024. ISSN 2503-2259. 4, 14

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*,

Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015. 3, 5, 11, 12

HSU, B.-M. Comparison of supervised classification models on textual data. *Mathematics*, MDPI AG, v. 8, n. 5, p. 851, maio 2020. ISSN 2227-7390.

ISLAM, M. R. *et al*. Machine learning enabled iot system for soil nutrients monitoring and crop recommendation. *Journal of Agriculture and Food Research*, Elsevier BV, v. 14, p. 100880, dez. 2023. ISSN 2666-1543.

JAPKOWICZ, N. *Assessment Metrics for Imbalanced Learning*. [S.l.]: Wiley, 2013. 187–206 p. 4, 9, 12

JUNIOR, G. d. B. V. *et al*. MÉtricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes. *Centro de Pesquisas Avançadas em Qualidade de Vida*, Revista CPAQV, v. 14, n. v14n2, p. 1, 2022. ISSN 2178-7514. 4,

KATARYA, R. *et al*. Impact of machine learning techniques in precision agriculture. In: *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*. [S.l.]: IEEE, 2020. LIAKOS, K. *et al*. Machine learning in agriculture: A review. *Sensors*, MDPI AG, v. 18, n. 8, p. 2674, ago. 2018. ISSN 1424-8220. 21

M., S. K.; MEGALINGAM, R. K. A survey on machine learning in agriculture - background work for an unmanned coconut tree harvester. In: *2019 Third InternationalConference on Inventive Systems and Control (ICISC)*. [S.l.]: IEEE, 2019. v. 9, p.433–437.

MARIAMMAL, G. *et al*. Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers. *IEEE Transactions on Computational Social Systems*, Institute of Electrical and Electronics Engineers (IEEE), v. 8, n. 5, p. 1132–1142, out. 2021. ISSN 2373-7476.

NAGARAJA, G. S. *et al*. Iot based smart agriculture management system. In: *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*. [S.l.]: IEEE, 2019.

P, K. M.; R, D. N. Crop prediction based on influencing parameters for different states in india- the data mining approach. In: *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. [S.l.]: IEEE, 2021. p. 1785–1791.

PRUSTY, S.; PATNAIK, S.; DASH, S. K. Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, Frontiers Media SA, v. 4, ago. 2022. ISSN 2673-3013.

RANJAN, P.; GARG, R.; RAI, J. K. Artificial intelligence applications in soil amp; crop management. In: *2022 IEEE Conference on Interdisciplinary Approaches in*

Technology and Management for Social Innovation (IATMSI). [S.l.]: IEEE, 2022. v. 8, p. 1–5.

RAO, J. S. *et al*. Ai, ar enabling on embedded systems for agricultural drones. In: *2022International Conference on Futuristic Technologies (INCOFT)*. [S.l.]: IEEE, 2022. p. 1–4.

SARAIVA, A. M. *et al*. A inteligência artificial na pesquisa agrícola. *Revista USP*, n. 141, p. 91-106, abr./jun. 2024.

SAVLA, D. V. *et al*. Virtual farmer: Real time crop prediction and automatic irrigation system. *In*: *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. [S.l.]: IEEE, 2020. v. 5, p. 1–5.

SIZAN, N. S. *et al*. Revolutionizing agriculture: An iot-driven ml-blockchain framework 5.0 for optimal crop prediction. In: *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*. [S.l.]: IEEE, 2023. p. 1–6. 10

SPEISER, J. L. *et al*. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, Elsevier BV, v. 134, p. 93–101, nov. 2019. ISSN 0957-4174.

TILMAN, D. *et al*. Agricultural sustainability and intensive production practices. *Nature*,Springer Science and Business Media LLC, v. 418, n. 6898, p. 671–677, ago. 2002. ISSN 1476-4687.

XIA, S.; YANG, Y. *A model-free feature selection technique of feature screening and random forest based recursive feature elimination*. [S.l.]: arXiv, 2023.

ZHANG, M. Comparing roc curves on multiclass classification for predicting quality of wine. *Worcester Polytechnic Institute*, 2021. Disponível em: https://digital.wpi.edu/concern/etds/ws859j70j?locale=en.