Modelagem Preditiva Para Determinação do Consumo de Água em Lavouras Utilizando Redes Neurais Artificiais e Random Forest Regressor

André Luiz Bandeli Júnior; Luiz Fernando Ferreira da Silva Neto; Ronaldo Lopes da Silva Filho

Faculdade de Engenharia Agrícola da Unicamp
Introdução à Mineração de Dados Aplicados à Agricultura

Projeto Final de Disciplina

RESUMO

A utilização de Deep Learning em problemas relacionados à agricultura têm crescido nos últimos anos, ganhando relevância enquanto ferramentas de predição por meio do aprendizado profundo. Neste sentido, o presente trabalho apresenta a implementação de uma Rede Neural Artificial (RNA) e um Modelo de Random Forest Regressor (RFR) para prever o uso de água com base em sete variáveis: tipo de cultura, área da fazenda (ha), tipo de irrigação, uso de fertilizantes (ton), uso de pesticidas (kg), produção total (ton) e tipo de solo. A base de dados foi escolhida com base em dois critérios: (i) relevância da preservação dos recursos hídricos e (ii) potencial de aplicação para maximizar a eficiência em sistemas irrigados. A metodologia inclui a implementação de um processo KDD para uma análise exploratória inicial e remoção de nulos e outliers. O dataset é dividido em 80% treino e 20% teste e geração de visualização gráfica com Matplot. O resultado demonstrou que, para a base inicial, ambos os modelos apresentaram dificuldades em capturar padrões, com R^2 negativo. Ao aumentar a base de dados, o Random Forest apresentou melhor desempenho (R^2 0,98) quando comparado a Rede Neural Artificial (R^2 0,80). Embora os resultados sejam relevantes, ressaltamos o critério experimental da base de dados aumentada, que possui limitações em contextos reais.

INTRODUÇÃO

A análise de dados relacionados ao consumo de água é fundamental para a agricultura, devido às diferentes aplicações, como o manejo e a otimização da irrigação, o planejamento de culturas e rotação, a gestão agrícola e o manejo de solos. Devido ao alto uso de água para a agricultura, aos altos custos de energia e à crescente preocupação mundial com os recursos hídricos, torna-se necessária a adoção de estratégias de manejo que possibilitem economia de água sem prejuízos à produtividade das culturas que demandam irrigação (MANTOVANI et al., 2013).

Nesse contexto, o uso de informações relacionadas à produtividade e à eficiência do uso da água são importantes não apenas para a análise econômica do sistema produtivo, mas também para auxiliar na economia do uso da água. Diante disso, é crescente o uso de abordagens baseadas em Machine Learning e Deep Learning para identificar padrões e realizar predições mais precisas que auxiliem na tomada de decisão no setor agrícola. O presente trabalho busca apresentar, portanto, uma análise descritiva de dados de irrigação e um modelo preditivo para valores de consumo de água, utilizando Redes Neurais Artificiais, como projeto final para a disciplina de Mineração de Dados Aplicados à Agricultura.

OBJETIVOS

O processo envolveu o download da base de dados de irrigação, seguido da conversão dos arquivos para o formato .xlsx. Em seguida, aplicou-se a metodologia de Descoberta de Conhecimento em Dados (KDD) para preparação e análise dos dados. Com os dados tratados, implementou-se um modelo de Rede Neural Artificial, cuja performance foi avaliada por meio da geração de gráficos e tabelas. Posteriormente, desenvolveu-se um script para realizar predições com base no modelo treinado. Por fim, os resultados obtidos foram apresentados e discutidos, destacando o desempenho e a aplicabilidade do modelo.

METODOLOGIA

A metodologia que será aplicada no estudo está descrita a seguir, correspondendo aos passos de download da base de dados, pré-processamento, aplicação de KDD, implementação da rede neural, gráficos, métricas de avaliação do modelo e implementação do preditor.

BASE DE DADOS

A escolha da base de dados sobre irrigação e consumo de água justifica-se pela sua relevância na análise da eficiência do uso hídrico em sistemas agrícolas irrigados, fundamental para a preservação dos recursos naturais. Ao escolher esta base, buscaremos realizar análises preditivas ao relacionar variáveis como tipo de irrigação, área cultivada, uso de insumos, produtividade da lavoura e características do solo com o volume de água utilizado. A tabela 1 mostra as variáveis selecionadas para o desenvolvimento do trabalho.

Tabela 1. Variáveis selecionadas para estudo. Exemplificação das primeiras linhas. Legenda: CT = Crop Type, FA = Farm Area, IT = Irrigation System, FU = Fertilizer Used, PU = Pesticide Used, ST = Soil Type e WU = Water Used.

Farm_ID	CT	FA (acres)	IT	FU (ton)	PU (tons)	Yield (tons)	ST	WU (m³)
F001	Cotton	329.4	Sprinkler	8.14	2.21	14.44	Loamy	76648.2
F002	Carrot	18.67	Manual	4.77	4.36	42.91	Peaty	68725.54
F003	Sugarcane	306.03	Flood	2.91	0.56	33.44	Silty	75538.56

ALGORITMO

O algoritmo e os dados utilizados estão disponíveis para consulta e execução dos experimentos[1]. O script desenvolvido analisa o consumo de água em áreas agrícolas, iniciando com a importação, limpeza e análise exploratória dos dados, considerando variáveis como tipo de cultura, solo, irrigação e estação do ano. Após o pré-processamento (codificação, divisão e padronização), são aplicados dois modelos preditivos: Random Forest Regressor e Rede Neural (MLPRegressor), ambos com validação cruzada e ajuste de hiperparâmetros. A avaliação é feita com RMSE, R^2 e MAE, comparando o desempenho e a generalização de cada abordagem.

PROCESSAMENTO DOS DADOS

A análise e o pré-processamento da base de dados envolverão as seguintes etapas: avaliação e tratamento de valores faltantes e nulos, remoção de outliers com base na análise dos quartis (utilizando o intervalo interquartil para identificação de valores extremos), e transformação dos dados. Além disso, foi aplicado um KDD, conforme FAYYAD e STOLORZ (1997), com o objetivo de extrair maiores informações da base para auxiliar a modelagem do modelo de RNA.

REDES NEURAIS ARTIFICIAIS

As Redes Neurais Artificiais (RNAs) são sistemas que mapeiam vetores de entrada em vetores de saída (BRAGA et al., 2007), inspirados em sistemas biológicos humanos, em particular tentando simular as conexões que ocorrem no cérebro humano (CICEK e OZTURK, 2021; SUKAMTO et al., 2023). Sua relação funcional é definida implicitamente por meio de camadas interconectadas e funções de ativação, visando minimizar o erro entre saídas previstas e observadas (CICEK e OZTURK, 2021).

A modelagem foi realizada com a biblioteca *scikit-learn* em Python, utilizando o *MLPRegressor*. Para otimização de hiperparâmetros, aplicou-se *GridSearch (G. S e S. BRINDHA, 2022)* com validação cruzada (5 *folds*), avaliada pela métrica *negative mean squared error* (convertida para escala positiva). Apesar da complexidade envolvida na otimização de redes neurais profundas, estudos recentes demonstram que redes super-parametrizadas, nas quais o número de parâmetros excede o número de amostras de dados, podem ser

[1] Script e base de dados utilizadas. Disponível em: <https://github.com/andre-bandeli/irrigation_predictor>

otimizadas até a otimalidade global por meio de métodos locais como o gradiente descendente, contrariando a expectativa tradicional de que mínimos locais seriam um obstáculo significativo (SOLTANOLKOTABI, JAVANMARD e LEE, 2019). O espaço de busca configurou-se conforme a Tabela 1.

Tabela 2. Hiperparâmetros utilizados no treinamento da Rede Neural Artificial.

Hiperparâmetro	Valores testados
hidden_layer_sizes	(50,), (100,), (100, 50), (150, 100, 50)
alpha	0,0001; 0,001; 0,01
activation	ReLU; Tanh
learning_rate	Constant; Adaptive

A RNA foi selecionada inicialmente para modelagem preditiva devido à sua capacidade de aprender padrões não lineares entre variáveis, particularmente relevante em problemas envolvendo propriedades físicas do solo e uso de água. Ressalta-se, porém, que a base de dados limitada pode comprometer a assertividade do modelo, induzindo underfitting ou overfitting, isto é, cenário em que o algoritmo não consegue extrair padrões suficientes ou em que o modelo memoriza dados em vez de generalizar padrões (BEJANI e GHATEE, 2019; RICE , WONG e KOLTER, 2020).

RANDOM FOREST

O algoritmo Random Forest (BREIMAN, 2001) é uma técnica de aprendizado de máquina baseada em um conjunto de árvores de decisão independentes e não correlacionadas, que, ao serem combinadas por meio do método de bagging, promovem maior capacidade de generalização do modelo (MRABET et al., 2022). O modelo foi implementado como alternativa potencial à RNA, também utilizando a biblioteca scikit-learn do Python. Para otimização de hiperparâmetros, também foi aplicada a técnica GridSearch com validação cruzada, avaliando o desempenho pela métrica negative mean_squared_error. O espaço de busca abrangeu quatro parâmetros principais:

Tabela 2. Hiperparâmetros utilizados no treinamento da Rede Neural Artificial.

Hiperparâmetro	Valores testados
n_estimators	100, 200, 300
max_depth	10, 20, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4

O modelo base foi instanciado com random state de 42 e n_jobs = -1. Após ajuste aos dados de treinamento, o melhor estimador foi selecionado conforme menor erro quadrático médio, seguido de reavaliação via validação cruzada para mensuração do desempenho (média e desvio padrão do erro).

DADOS SINTÉTICOS

Para aprimorar a capacidade de predição do modelo, a base de dados original foi expandida através da geração de novos dados. Este processo foi executado utilizando a técnica de bootstrap, que consiste na reamostragem com reposição do conjunto de dados existente. Esta abordagem permitiu a criação de um volume adicional de registros que replicam as distribuições estatísticas e relações inerentes aos dados da base inicial. A inclusão desses dados sintéticos foi fundamental para superar as limitações impostas por um tamanho amostral restrito (50 linhas cada variável), contribuindo para a validação dos modelos propostos. No entanto, ressaltamos que esta prática, embora útil para fins experimentais, possui vulnerabilidades em experimentos com dados reais.

RESULTADOS

Os resultados iniciais demonstraram problemas e desafios na aplicação de modelagem de redes neurais artificiais e, também, na aplicação de árvores de decisão (Random Forest). Através da distribuição dos dados (Figura 1) é possível analisar os resultados exploratórios iniciais.

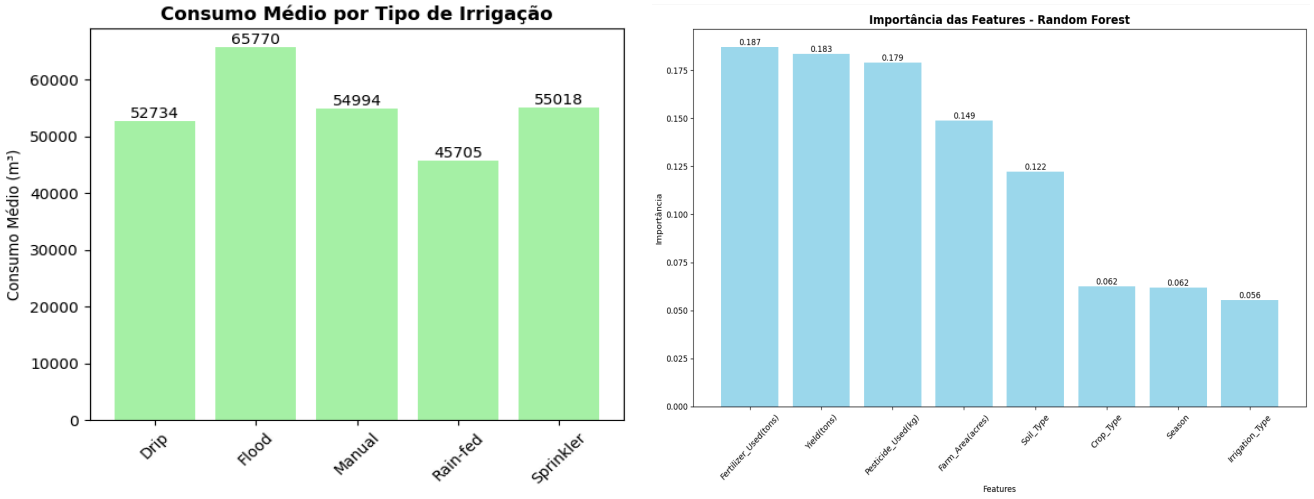

Figura 1. Consumo Médio por Tipo de Irreigação (esqueda) e Importância das Features (direita).

A matriz de correlação (Figura 2) mostra as relações lineares relevantes entre as variáveis preditoras e o consumo hídrico (Water_Usage). Destacam-se correlações moderadas com Yield (r = 0,18), Soil_Type (r = -0,21) e Season (r = -0,18), sugerindo que produtividade, tipo de solo e sazonalidade influenciam a demanda por água. As demais variáveis apresentaram correlações fracas, indicando influência limitada na variável alvo.

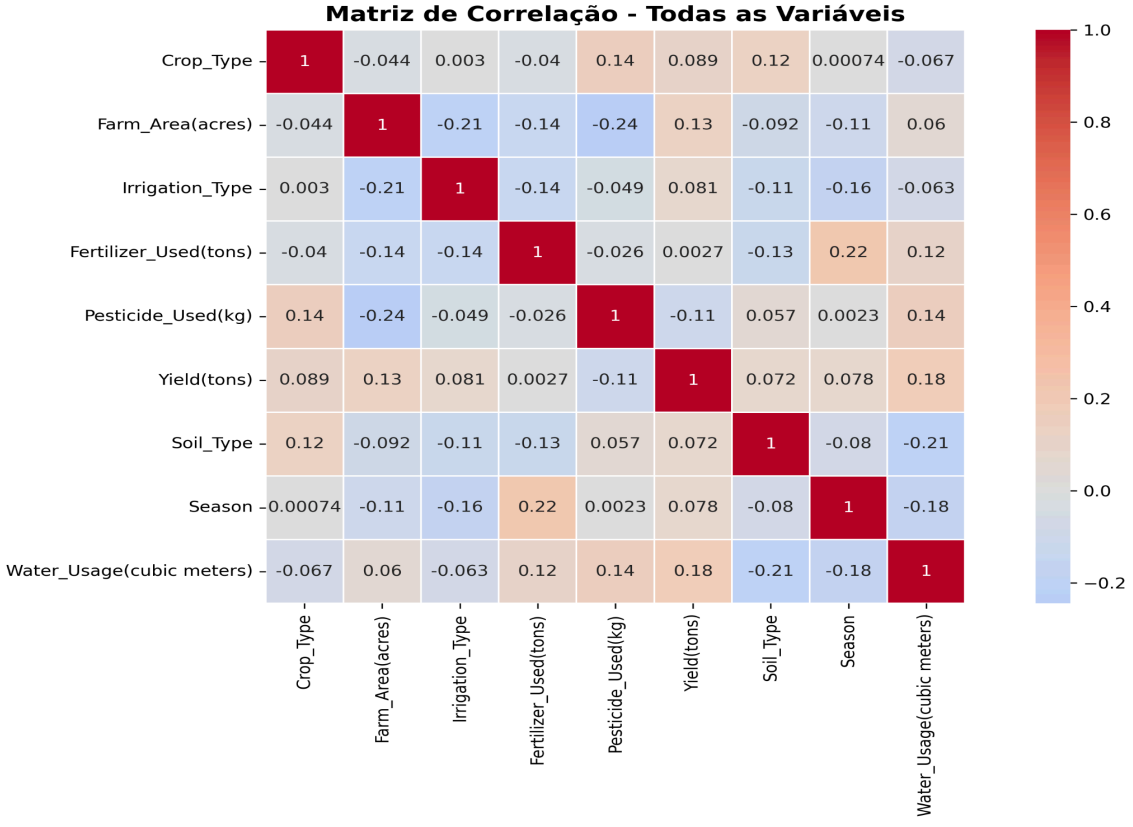

Figura 2. Matriz de Correlação entre as variáveis da base de dados.

A análise gráfica da Figura 3, que ilustra o consumo de água por tipo de cultura e a distribuição do consumo, além da presença de outliers e padrões não lineares que justificam a escolha de técnicas de aprendizado de máquina para esse tipo de problema, em detrimento de regressões lineares simples. Culturas como a cana-de-açúcar e o algodão apresentaram elevados volumes médios de uso de água, fato que pode estar relacionado tanto à área cultivada quanto ao tipo de solo e método de irrigação empregado.

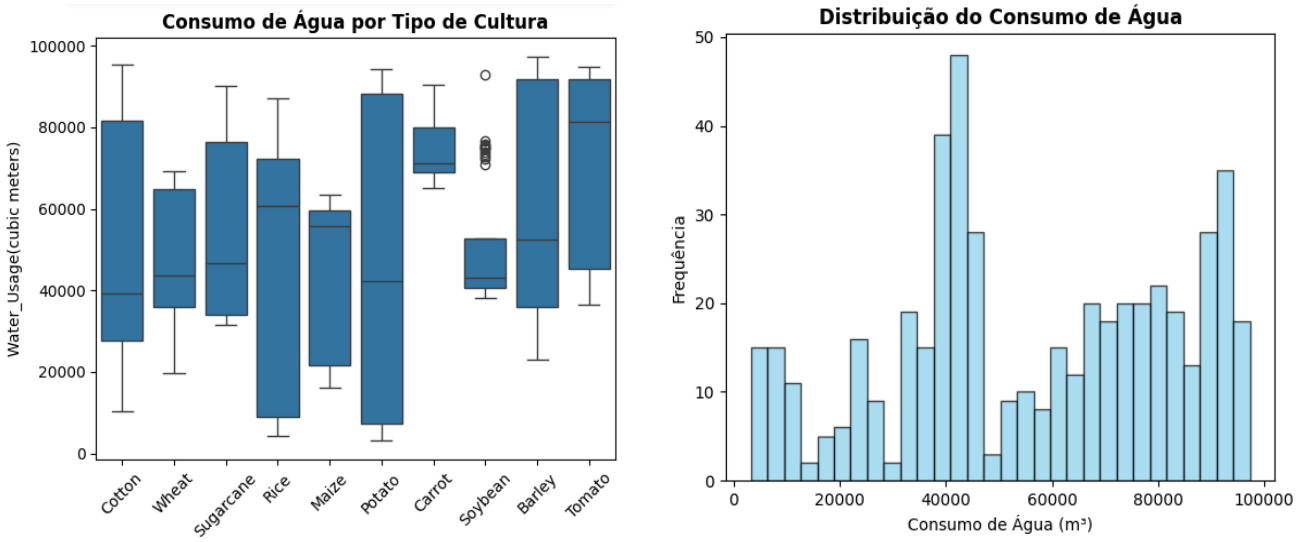

Figura 3. Consumo de Água por Tipo de Cultura (esquerda) e Distribuição do Consumo de Água (direita). .

Os valores das métricas R^2 demonstraram baixo desempenho. Os erros médios absolutos (MAE) e os erros quadráticos médios (MSE e RMSE) também apresentaram valores expressivamente altos, indicando uma alta divergência entre os valores reais e preditos. Esses resultados podem ser atribuídos à baixa quantidade de dados na base original, fator que compromete a generalização e o aprendizado dos modelos, conforme indicado na tabela 3.

Tabela 3. Resultados dos modelos utilizando a base de irrigação inicial.

Modelo	MSE	RMSE	MAE	R^2
Random Forest	9.517.950.059.921	308.511.751	250.370.269	-1.963
Rede Neural	9.335.039.905.310	305.532.975	240.249.707	-1.733

Diante disso, prosseguimos com a geração de dados sintéticos utilizando o método de bootstrap, expandindo a base de forma a manter as características estatísticas dos dados originais. A Tabela 4 apresenta os resultados após o aumento de dados na base inicial. Nota-se uma melhora substancial na performance dos modelos. O Random Forest obteve um R^2 de 0,98, enquanto a Rede Neural alcançou 0,80.

Tabela 4. Resultados dos modelos utilizando a base de irrigação aumentada.

Modelo	MSE	RMSE	MAE	R^2
Random Forest	12172103.6864	3488.8542	2343.1955	0.9817
Rede Neural	126580423.2869	11250.7966	8774.3947	0.8097

Os gráficos de dispersão apresentados na Figura 4 corroboram esses resultados: observa-se uma aproximação mais evidente entre os valores reais e preditos, principalmente no modelo Random Forest, o que reforça sua capacidade de generalização e menor sensibilidade ao overfitting, característica esperada conforme Breiman (2001).

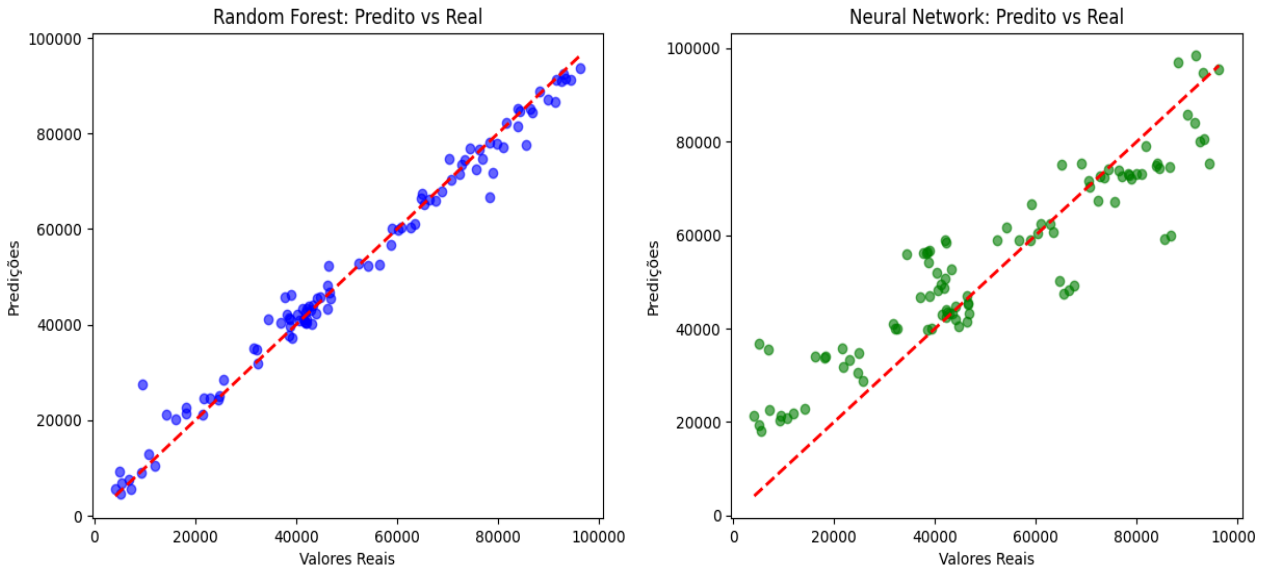

Figura 4. Gráficos de Distribuição dos Valores Reais versus Preditos. Random Forest (esquerda) e Rede Neural (direita).

Os melhores hiperparâmetros obtidos por meio da técnica de GridSearch foram, para a Rede Neural Artificial (RNA): activation='relu', alpha=0.0001, hidden_layer_sizes=(150, 100, 50) e learning_rate='constant', resultando em um CV Score de 212.071.568,99 ± 79.965.785,81. Para o modelo Random Forest Regressor (RFR), os melhores parâmetros encontrados foram: max_depth=10, min_samples_leaf=1, min_samples_split=2 e n_estimators=200, com um CV Score de 24.145.757,77 ± 24.420.981,74.

Por fim, os resultados sugerem que, embora as Redes Neurais sejam eficazes em problemas com alta não linearidade, o modelo Random Forest apresentou maior robustez e estabilidade, especialmente frente à escassez de dados, corroborando estudos anteriores que recomendam seu uso em cenários com conjuntos de dados reduzidos ou ruidosos (BREIMAN, 2001; CHEN; ISLAM, 2019)

CONCLUSÕES

[1] A seleção da base de dados voltada ao consumo de água na agricultura mostrou-se relevante diante do contexto estudado, visando estudos técnicos para melhorar a conservação do uso de água em lavouras. A aplicação - com dados reais - pode apoiar decisões estratégicas relacionadas ao uso eficiente da água em sistemas irrigados, contribuindo para o desenvolvimento de uma agricultura mais sustentável.

[2] O processo de pré-processamento e análise exploratória dos dados foi essencial para a identificação de padrões e inconsistências na base. A remoção de outliers e a padronização dos dados forneceram maior confiabilidade na modelagem. A aplicação do KDD, aliada à geração de dados sintéticos, foi fundamental para mitigar os efeitos da limitação amostral.

[3] A aplicação da Rede Neural Artificial demonstrou potencial para capturar relações não lineares entre as variáveis, especialmente em contextos onde há alta complexidade nos dados. No entanto, os resultados obtidos com a base original evidenciaram a sensibilidade do modelo à escassez de dados, resultando em baixa acurácia e alto erro preditivo.

[4] O modelo Random Forest apresentou melhor desempenho em todos os indicadores avaliados, tanto com a base original quanto com a base aumentada. No entanto, com a base inicial, também apresentou R^2 negativo, embora com desempenho superior a Rede Neural. Destaca-se a menor propensão ao overfitting, sendo mais adequado para o cenário de dados reduzidos.

[5] A estratégia de ampliação da base com dados sintéticos, via bootstrap, impactou positivamente o desempenho dos modelos, principalmente da Random Forest, que alcançou valores elevados de R^2. Ainda que não substitua dados reais, a abordagem foi eficaz como solução experimental para compensar a limitação do conjunto original.

[6] Apesar da expectativa quanto ao desempenho das RNAs, o modelo baseado em árvores de decisão se mostrou superior no presente estudo, especialmente em termos de estabilidade e capacidade de generalização. Os resultados destacam a importância de considerar o volume e a qualidade dos dados ao escolher a abordagem de aprendizado de máquina mais adequada.

[7] Conclui-se que o uso de modelos de aprendizado de máquina pode contribuir significativamente para a previsão do consumo hídrico na agricultura. A integração de técnicas de pré-processamento, modelagem e validação possibilita aplicações práticas na gestão de recursos, desde que acompanhadas por bases de dados consistentes e representativas da realidade agrícola. Estudos futuros podem ampliar essa análise incorporando dados climáticos, temporais e geográficos para aumentar ainda mais a precisão dos modelos preditivos.

REFERÊNCIAS BIBLIOGRÁFICAS

BEJANI, Mohammad Mahdi; GHATEE, Mehdi. **Regularized deep networks in intelligent transportation systems: a taxonomy and a case study**. Artificial Intelligence Review, 2021. Disponível em: https://doi.org/10.1007/s10462-021-09975-1

BREIMAN, Leo. **Random forests**. Machine Learning, v. 45, p. 5–32, 2001.

CICEK, Zeynep Idil Erzurum; OZTURK, Zehra Kamisli. **Optimizing the artificial neural network parameters using a biased random key genetic algorithm for time series forecasting**. Applied Soft Computing Journal, v. 102, 107091, 2021. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1568494621000170

EL MRABET, Zakaria et al. **Random forest regressor-based approach for detecting fault location and duration in power systems**. Sensors, v. 22, n. 2, p. 458, 2022. DOI: https://doi.org/10.3390/s22020458.

FAYYAD, Usama; STOLORZ, Paul. **Data mining and KDD: promise and challenges**. Future Generation Computer Systems, v. 13, p. 99–115, 1997.

G., S.; BRINDHA, S. **Hyperparameters optimization using Gridsearch Cross Validation Method for machine learning models in predicting diabetes mellitus risk**. In: INTERNATIONAL CONFERENCE ON COMMUNICATION, COMPUTING AND INTERNET OF THINGS (IC3IoT), 2022, Chennai, Índia. Anais [...]. IEEE, 2022.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. **Deep learning**. Nature, v. 521, p. 436–444, 2015. DOI: https://doi.org/10.1038/nature14539.

MANTOVANI, E. C. et al. **Eficiência no uso da água de duas cultivares de batata-doce em resposta a diferentes lâminas de irrigação**. Horticultura Brasileira, v. 31, n. 4, p. 602–606, 2013.

RICE, Leslie; WONG, Eric; KOLTER, J. Zico. **Overfitting in adversarially robust deep learning**. 2020. Disponível em: https://doi.org/10.48550/arXiv.2002.11569

SOLTANOLKOTABI, Mahdi; JAVANMARD, Adel; LEE, Jason D. **Theoretical insights into the optimization landscape of over-parameterized shallow neural networks**. IEEE Transactions on Information Theory, v. 65, n. 2, p. 742–769, fev. 2019. DOI: https://doi.org/10.1109/TIT.2018.2869182.

SUKAMTO; HADIYANTO; KURNIANINGSIH. **KNN optimization using Grid Search algorithm for preeclampsia imbalance class.** E3S Web of Conferences, v. 448, 02057, 2023. DOI: https://doi.org/10.1051/e3sconf/202344802057.