



UNIVERSIDADE FEDERAL DE PERNAMBUCO  
CENTRO ACADÊMICO DO AGRESTE  
NÚCLEO DE TECNOLOGIA  
CURSO DE GRADUAÇÃO EM ENGENHARIA CIVIL

JEAN FIRMINO CARDOSO

**ESTIMATIVA FUTURA DE VARIÁVEIS HIDROLÓGICAS UTILIZANDO TÉCNICAS  
DE APRENDIZAGEM DE MÁQUINA: UMA ABORDAGEM COMPARATIVA**

Caruaru  
2024

JEAN FIRMINO CARDOSO

**ESTIMATIVA FUTURA DE VARIÁVEIS HIDROLÓGICAS UTILIZANDO TÉCNICAS  
DE APRENDIZAGEM DE MÁQUINA: UMA ABORDAGEM COMPARATIVA**

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Engenharia Civil do Campus Agreste da Universidade Federal de Pernambuco – UFPE, na modalidade de artigo científico, como requisito parcial para obtenção do grau de Bacharel em Engenharia Civil.

**Área de concentração:** Hidrologia

**Orientador:** Prof. D.Sc. Saulo de Tarso Marques Bezerra

Caruaru

2024

*"Diz-se que o sabor das lágrimas é diferente, pois as lágrimas de alegria e tristeza são doces enquanto as lágrimas de frustração e raiva são salgadas! Existe uma pessoa no mundo que chora porque quer chorar? Quero viver minha vida sem arrependimentos, então só quero chorar doces lágrimas de alegria!"*  
*- Eiichiro Oda, One Piece, Vol 108.*

## **AGRADECIMENTOS**

Minha gratidão à família, em particular aos meus pais, Sônia Firmino e Josué Cardoso, por sempre me encorajarem e fornecerem todo tipo de apoio necessário para que eu pudesse me dedicar aos meus mais grandiosos sonhos. Eles desempenharam um papel fundamental na realização deste momento e de muitos outros em minha vida. Amo muito vocês dois!

Desejo expressar minha profunda gratidão ao meu orientador de TCC, o professor Saulo de Tarso, pelo constante apoio, orientação e paciência durante toda a minha pesquisa. Suas observações perspicazes desempenharam um papel fundamental na formatação e melhoria da qualidade do meu trabalho.

Agradeço aos professores Gilson Lima e Andreia Azevedo por aceitarem o convite em serem os membros da banca examinadora, assim contribuindo para o aprimoramento do meu trabalho de conclusão de curso.

Um agradecimento para Artur Coutinho, Lucas Ravellys, Severino Martins e Antônio Antonino por estarem ao meu lado durante toda a minha jornada acadêmica, oferecendo apoio, incentivo e motivação desde o início.

Deixo meus votos de agradecimento a todos os professores do curso de Engenharia Civil do campus do Agreste, que compartilharam um pouco do seu conhecimento técnico, humano e espiritual durante as aulas e as conversas de corredor. Meu mais profundo respeito e carinho por todos vocês.

Também quero expressar minha gratidão aos meus queridos amigos Allana Katyelenn, André Inácio, Antônio Nascimento, Arnaldo Junior, Breno Vilela, Bruno Henrique, Carina Lessa, Caio Vinicius, Daniel Queiroz, Danylo Leal, Diogo Nobre, Douglas Rafael, Elyaquim Torres, Erica Rafaela, Gabriel Portela, Higor Gustavo, Iago Leopoldo, Ingrid Karoline, Isabelle Sales, Ítalo Silvio, Isabela Freitas, Ivan Nunes, Jhuan Miguel, José Ângelo, Lucas Amâncio, Marcos Vinicius, Maria do Carmo, Maria Rita, Mariana Pinheiro, Pedro Barbosa, Pedro Henrique, Phelipe Monteiro, Rodrigo Gonçalves, Soffia Valéria, Tiago Iluminato, Túlio Henrique, Urgel Júnior, Un Hee Martha, Vinicius de Moraes e a todos os amigos que estiveram muito presentes em minha vida, tornando a jornada acadêmica mais feliz e menos desafiadora, estão eternamente nas minhas melhores memórias e lembranças dessa fase da vida.

Manifesto minha sincera apreciação ao INMET, que disponibilizou os dados essenciais que tornaram minha pesquisa possível. Agradeço às organizações Pórtis

Júnior, Beltrão Engenharia, PDCA Engenharia, CNPq, Facepe, Méliuz, Borborema Finance, Pescaria Universitária e aos projetos MAI-DAI e ao ONDACBC por desempenharem um papel significativo em minha formação pessoal e profissional. Essas experiências ampliaram minha perspectiva, proporcionando oportunidades para participar de projetos importantes e adquirir habilidades práticas e conhecimentos teóricos que serão inestimáveis para minha carreira e vida.

Aos amigos de infância, como Felipe Monteiro, Luana Ribeiro, Lucas José, William Sabino, Matheus Santos, Márcio Ladyson, João Lucas, Davyd Edvaldo e Mário Mota, que estiveram ao meu lado, oferecendo incentivo e apoio nas várias fases da minha vida, agradeço de coração.

Quero expressar minha mais sincera gratidão a todos que compartilharam seu tempo e suas experiências em me auxiliar de alguma forma. Sem a generosidade de vocês em compartilhar seus sonhos, experiências e pensamentos, meu progresso como pessoa não teria sido possível. Não importa o quão difícil ou triste seja o seu momento presente, em algum lugar, alguém sempre estará desejando a sua felicidade, esse alguém você já entendeu quem é.

## LISTA DE ABREVIATURAS E SIGLAS

|         |  |
|---------|--|
| UFPE    | Universidade Federal de Pernambuco                                       |
| CAA     | Campus do Agreste  |
| ANA     | Agência Nacional de Águas e Saneamento Básico                            |
| INMET   | Instituto Nacional de Meteorologia                                       |
| ONDACBC | Observatório Nacional da Dinâmica da Água e do Carbono no Bioma Caatinga |
| NSE     | Coefficiente de Eficiência de Nash-Sutcliffe                             |
| WI      | Índice de Willmott   |
| c       | índice de desempenho do modelo   |
| KGE     | Índice de Eficiência Kling-Gupta   |
| SM      | Somatório das Métricas   |
| RMSE    | Erro Quadrático Médio da Raiz  |
| TDR     | <i>Time Domain Reflectometry</i>   |
| ANN     | <i>Artificial Neural Networks</i>  |
| SVM     | <i>Support Vector Machines</i>   |
| DT      | <i>Decision Trees</i>  |
| RF      | <i>Random Forests</i>  |
| GB      | <i>Gradient Boosting</i>   |
| R       | Coefficiente de correlação de Pearson                                    |

## LISTA DE SÍMBOLOS

|                   |  |
|-------------------|--|
| $m_a$             | Massa de água no solo                          |
| $V_a$             | Volume de água no solo                         |
| $m_s$             | Massa no solo seco                             |
| $V_t$             | Volume total do solo                           |
| $\theta_v$        | Umidade volumétrica no solo                    |
| $\theta_r$        | Umidade residual                               |
| $\theta_s$        | Umidade saturada                               |
| $E_i'$            | Informação temporal normalizada para o tempo i |
| $E_i$             | Informação temporal no tempo i                 |
| $média(E \ln)$    | Média do $\ln(E_i+1)$ de treino                |
| $S \ln$           | Desvio padrão do $\ln(E_i)$ de treino          |
| $y_i$             | Valor medido                                   |
| $\hat{y}_i$       | Valor estimado                                 |
| $mean(y)$         | Média dos valores medidos                      |
| $mean(\hat{y}_i)$ | Média dos valores estimado                     |
| $N$               | Número de registros medidos                    |
| $\beta$           | Razão de viés                                  |
| $\gamma$          | Relação de variabilidade                       |
| $CV$              | Coefficiente de variação dos valores           |

estimados e medidos

$\sigma$

Desvio padrão dos valores estimados e medidos

$\mu$

Média dos valores estimados e medidos

$\hat{y}_i$

Valor estimado

$E_i$

Informação temporal no tempo i

$E'_i$

Informação temporal normalizada para o tempo i

$y_i$

Valor medido

## **Estimativa futura de variáveis hidrológicas utilizando técnicas de aprendizagem de máquina: uma abordagem comparativa**

### **Future estimation of hydrological variables using machine learning techniques: a comparative approach**

**Jean Firmino Cardoso<sup>1</sup>**

---

#### **RESUMO**

A previsão e estimativa de variáveis hidrológicas desempenham um papel crucial na gestão sustentável dos recursos hídricos, sendo essenciais para o planejamento e tomada de decisões em diversas áreas, como agricultura, engenharia civil, controle de enchentes e geração de energia. No entanto, os processos hidrológicos apresentam características complexas e não lineares, o que representa um desafio para a modelagem e previsão precisa dessas variáveis. Diante desse desafio, as técnicas de aprendizagem de máquina têm se destacado como uma abordagem promissora para a estimativa futura da vazão de rios e outras variáveis hidrológicas. Neste trabalho, foi proposta uma abordagem comparativa de técnicas de aprendizagem de máquina aplicadas a conjuntos de dados hidrológicos. O objetivo principal foi analisar e comparar diferentes modelos de aprendizagem de máquina para identificar qual técnica ou combinação de técnicas apresenta o melhor desempenho na previsão de variáveis hidrológicas. Utilizou-se técnicas de aprendizagem de máquina, como máquinas de suporte vetorial, árvore de decisão, florestas aleatórias, redes neurais artificiais e gradiente *boosting*, selecionadas por sua capacidade de lidar com a complexidade e não linearidade dos processos hidrológicos. Os resultados obtidos apresentaram uma capacidade de previsibilidade boa, variando de 50 a 70% sua eficiência tendo como base o conjunto de métricas selecionadas, sendo o modelo SVM com o pior desempenho de todos, pois variou de 30 a 40% de eficácia nas combinações de valores de entrada e saída que tiveram uma boa média dos resultados para os outros algoritmos. A análise comparativa dos resultados permitiu identificar padrões e relações entre variáveis e configurações iniciais dos algoritmos, contribuindo para uma melhor compreensão dos processos hidrológicos e sua previsibilidade.

<sup>1</sup>Graduando em Engenharia Civil pela Universidade Federal de Pernambuco. E-mail: jean.firmino@ufpe.br

**Palavras-chave:** estimativa hidrológica; aprendizagem de máquina; vazão de rios; recursos hídricos; modelos computacionais.

---

## **ABSTRACT**

The forecasting and estimation of hydrological variables play a crucial role in the sustainable management of water resources, being essential for planning and decision-making in various areas such as agriculture, civil engineering, flood control, and energy generation. However, hydrological processes exhibit complex and nonlinear characteristics, posing a challenge for the accurate modeling and prediction of these variables. Faced with this challenge, machine learning techniques have emerged as a promising approach for the future estimation of river flow and other hydrological variables. In this work, a comparative approach of machine learning techniques applied to hydrological datasets was proposed. The main objective was to analyze and compare different machine learning models to identify which technique or combination of techniques presents the best performance in predicting hydrological variables. Machine learning techniques such as support vector machines, decision trees, random forests, artificial neural networks, and gradient boosting were used, selected for their ability to handle the complexity and nonlinearity of hydrological processes. The results obtained showed good predictability, ranging from 50 to 70% efficiency based on the selected set of metrics, with the SVM model performing the worst of all, ranging from 30 to 40% effectiveness in input-output combinations that had a good average of results for the other algorithms. The comparative analysis of the results allowed for the identification of patterns and relationships between variables and initial algorithm configurations, contributing to a better understanding of hydrological processes and their predictability.

**Keywords:** hydrological estimation; machine learning; river flow; water resources; computational models.

---

**DATA DE APROVAÇÃO:** 12 de Março de 2024

---

## 1 INTRODUÇÃO

A previsão e estimativa de variáveis hidrológicas desempenham um papel crucial na gestão sustentável dos recursos hídricos. Essas informações são essenciais para o planejamento e tomada de decisões em diversas áreas, como gestão de recursos hídricos, agricultura, engenharia civil, controle de enchentes e geração de energia. No entanto, as características complexas e não lineares dos processos hidrológicos representam um desafio significativo para a modelagem e previsão precisa dessas variáveis.

A intrincada interação entre fatores naturais, como precipitação, evaporação, infiltração e escoamento, dificulta a criação de modelos robustos. Além disso, a influência das atividades humanas, como urbanização e desmatamento, complica ainda mais a equação. A variabilidade espacial e temporal desses fenômenos exige abordagens sofisticadas para capturar com precisão as mudanças ao longo do tempo.

A relevância desse estudo é ressaltada diante da crescente demanda por estimativas precisas de variáveis hidrológicas. O avanço nessa área pode oferecer uma contribuição significativa para aprimorar a gestão dos recursos hídricos, permitindo a alocação mais eficiente desses recursos, a redução dos riscos de enchentes e a otimização do uso da água em contextos agrícolas, industriais e domésticos.

As implicações práticas das previsões hidrológicas precisas são vastas, incluindo um melhor planejamento para a agricultura, controle de poluição, operações de usinas hidrelétricas e medidas preventivas contra desastres naturais (Liakos *et al.*, 2018). Além disso, de acordo com Muhammad *et al.* (2020), tais estimativas respaldam a tomada de decisões informadas em relação a projetos de desenvolvimento sustentável, garantindo a proteção dos ecossistemas aquáticos e a disponibilidade contínua de recursos para as gerações futuras.

Diversos estudos têm sido realizados no campo da estimativa das variáveis hidrológicas utilizando técnicas de aprendizagem de máquina. Por exemplo, Dayal *et al.* (2021) propuseram uma abordagem baseada em fluxos de água obtidos por satélite e técnicas de aprendizado de máquina para estimar a vazão de rios em regiões monçônicas da Índia. Huang *et al.* (2019) combinaram múltiplos modelos baseados em dados para previsões de longo prazo de escoamento mensal,

utilizando a média bayesiana dos modelos. Liang *et al.* (2018) integraram o gerador de precipitação aleatória de florestas aleatórias ao SWAT para realizar previsões de longo prazo de vazão em uma bacia de estudo específica. Prieto *et al.* (2019) utilizaram técnicas de florestas aleatórias para prever o fluxo em bacias não monitoradas, com a aplicação de testes estatísticos para avaliar a adequação dos resultados.

Diante desse desafio, técnicas de aprendizagem de máquina têm sido amplamente exploradas como uma abordagem promissora para a estimativa futura da vazão de rios e outras variáveis hidrológicas (Nachappa *et al.*, 2020). O aprendizado de máquina envolve a construção e o treinamento de modelos computacionais capazes de identificar padrões e relacionamentos em conjuntos de dados históricos (Braz *et al.*, 2023). Esses modelos são então utilizados para fazer previsões e estimativas em novos dados.

No cenário contemporâneo da análise de dados e do aprendizado de máquina, a *Support Vector Machine Regression* (SVMR) emerge como uma abordagem robusta e versátil para a modelagem de problemas de regressão. A SVMR é uma extensão natural da *Support Vector Machine* (SVM), uma técnica consagrada na literatura científica por sua eficácia na resolução de problemas de classificação (Lal; Datta, 2019). Ao adaptar os princípios da SVM para a tarefa de regressão, a SVMR oferece uma solução poderosa para a estimativa de valores contínuos, demonstrando sua aplicabilidade em uma ampla gama de campos.

A fundamentação da SVMR repousa sobre o conceito central de encontrar um hiperplano de regressão ótimo que melhor se ajusta aos dados observados (Lal; Datta, 2019). Diferentemente das abordagens de regressão tradicionais, que buscam minimizar as diferenças entre os valores observados e preditos, a SVMR se concentra na maximização da margem entre os pontos de dados e o hiperplano de regressão. Essa margem descreve a zona onde os pontos de dados podem se encontrar enquanto ainda estão dentro da margem de erro permitida, resultando em uma abordagem que privilegia a generalização e a robustez do modelo.

A operacionalização da SVMR envolve a transformação dos dados originais em um espaço de maior dimensionalidade, onde a busca pelo hiperplano de regressão ótimo é realizada (Kamir *et al.*, 2019). Esse processo é facilitado pela utilização de funções de kernel, que permitem a representação eficiente dos dados em espaços não-lineares. Assim como na SVM de classificação, a escolha do kernel apropriado

na SVMR é essencial para capturar relações complexas entre as variáveis independentes e a variável dependente.

Uma característica distintiva da SVMR é sua capacidade de lidar com outliers e não linearidades de maneira eficaz, resultando em modelos de regressão mais flexíveis e adaptáveis a diferentes contextos (Dalla Torre *et al.*, 2024). Além disso, a SVMR oferece a possibilidade de lidar com problemas de alta dimensionalidade, onde a relação entre as variáveis pode ser intrincada e difícil de ser modelada por métodos tradicionais (Dalla Torre *et al.*, 2024).

A regressão por Árvores de Decisão (*Decision Tree Regression*) é uma técnica proeminente no campo da análise de dados e aprendizagem de máquina, que visa abordar problemas de previsão de valores contínuos. Com base no conceito de estrutura de árvore hierárquica, essa abordagem oferece uma maneira intuitiva e poderosa de modelar relações complexas entre variáveis independentes e a variável de saída (Latif; Ahmed, 2023).

A Árvore de Decisão é uma representação gráfica que divide os dados em segmentos menores, estruturados como um fluxograma de decisões. Cada ramo da árvore representa um critério de divisão baseado em uma variável preditora, e os nós finais da árvore contêm as previsões de valores contínuos. Esse processo de divisão e segmentação dos dados permite a criação de modelos de regressão não-linear, capturando nuances e interações entre as variáveis que podem ser difíceis de serem expressas por métodos de regressão linear tradicionais (Latif; Ahmed, 2023). A construção de uma Árvore de Decisão de Regressão envolve a identificação das melhores divisões para maximizar a redução da variabilidade na variável de saída. A métrica de impureza, como a soma dos quadrados dos erros (SSE), é frequentemente usada para avaliar a qualidade das divisões e orientar a escolha das características mais relevantes. Esse processo iterativo de divisão e avaliação é repetido até que um critério de parada seja alcançado, resultando em uma estrutura de árvore que melhor se adapta aos dados de treinamento.

Uma das principais vantagens da Regressão por Árvores de Decisão é sua capacidade de capturar relações não-lineares e interações complexas entre variáveis, tornando-se especialmente útil em cenários onde as relações entre os dados são intrincadas e difíceis de serem modeladas por métodos lineares (Dumitrescu *et al.*, 2022). Além disso, as Árvores de Decisão são menos sensíveis a outliers e podem lidar eficazmente com variáveis categóricas e numéricas,

tornando-as versáteis para uma variedade de problemas.

A Regressão por Floresta Aleatória (*Random Forest Regression*) é uma abordagem avançada de aprendizado de máquina que tem ganhado destaque no campo da análise de dados e previsão de valores contínuos. Ela representa uma extensão da técnica de Árvores de Decisão, visando superar algumas de suas limitações e aprimorar a precisão das previsões por meio da agregação de múltiplas árvores (Chen *et al.*, 2020).

De acordo com Chen *et al.* (2020), a Floresta Aleatória opera a partir do princípio de construir várias Árvores de Decisão independentes e combiná-las para formar um modelo robusto e geralmente mais preciso. Cada árvore é construída com base em uma amostra aleatória dos dados de treinamento, além de usar uma seleção aleatória de variáveis preditoras em cada divisão. Esse processo de aleatorização reduz a tendência das árvores individuais de se ajustarem demais aos dados de treinamento, resultando em um modelo mais estável e com menor probabilidade de *overfitting*.

A combinação das previsões das diferentes árvores na Floresta Aleatória é realizada por meio de uma média ou votação ponderada, resultando em uma previsão final mais precisa e robusta (Chen *et al.*, 2020). Além disso, a técnica permite a avaliação da importância relativa de cada variável na previsão, fornecendo insights sobre quais características têm maior impacto nas saídas do modelo.

A Regressão por Floresta Aleatória é amplamente utilizada em uma variedade de domínios, devido à sua capacidade de lidar com conjuntos de dados complexos, heterogêneos e de alta dimensionalidade. Esta é particularmente eficaz em cenários em que a linearidade das relações entre as variáveis é limitada e as interações entre os preditores são difíceis de serem modeladas por técnicas tradicionais de regressão linear (Antoniadis *et al.*, 2021).

A regressão por *Gradient Boosting* é uma técnica avançada de aprendizado de máquina que tem se destacado como uma poderosa abordagem para a previsão precisa de valores contínuos em diversos domínios. Ela faz parte da família de algoritmos de *ensemble*, que consiste na combinação de múltiplos modelos individuais para melhorar a performance e a generalização do modelo final. O *Gradient Boosting* é particularmente eficaz em cenários onde a complexidade das relações entre as variáveis preditoras e a variável resposta requer uma modelagem mais sofisticada (Xu *et al.*, 2023).

De acordo com Xu *et al.* (2023), o princípio central do *Gradient Boosting* é criar um modelo forte a partir da combinação de vários modelos fracos, como árvores de decisão rasas. O algoritmo opera de maneira sequencial, onde cada nova árvore é ajustada para corrigir os erros residuais deixados pelas árvores anteriores. Isso é alcançado ajustando os pesos das observações para dar mais importância aos casos mais difíceis de serem previstos. Em cada iteração do *Gradient Boosting*, o algoritmo busca encontrar a direção no espaço de parâmetros que mais reduz o erro residual. Isso é feito através do cálculo dos gradientes da função de perda em relação aos parâmetros do modelo. A adição das árvores subsequentes é feita de forma que minimiza o erro residual, resultando em uma aproximação gradual e refinada da função alvo. Esta técnica oferece várias vantagens, como a capacidade de modelar relações não-lineares, lidar com valores ausentes de forma eficaz e capturar interações complexas entre as variáveis. Além disso, a técnica fornece uma estimativa da importância das variáveis, auxiliando na interpretação dos resultados.

As Redes Neurais Artificiais (RNA) têm se estabelecido como uma das abordagens mais poderosas e versáteis no campo do aprendizado de máquina e análise de dados, sendo amplamente utilizadas para modelar relações complexas entre variáveis preditoras e variáveis de resposta contínuas. Estas são modelos computacionais inspirados na estrutura e funcionamento do cérebro humano, sendo compostas por camadas de neurônios interconectados, cada um realizando operações matemáticas em entradas ponderadas para produzir uma saída. As camadas intermediárias, conhecidas como camadas ocultas, permitem a transformação não-linear dos dados, tornando as RNA capazes de representar funções altamente complexas. A flexibilidade inerente às RNA permite a captura de padrões não-lineares e interações sutis presentes nos dados, tornando-as uma ferramenta valiosa em diversas aplicações.

De acordo com Alaneme e Mbadike (2019), na regressão por RNA, o objetivo é estimar um relacionamento funcional entre as variáveis preditoras e a variável de resposta contínua. O processo de treinamento envolve a otimização dos pesos e viés das conexões entre os neurônios, a fim de minimizar a diferença entre as previsões da rede e os valores reais da variável de resposta. Esse processo é realizado por meio de algoritmos de otimização, como o Gradiente Descendente, que ajustam gradualmente os parâmetros da rede para melhor se ajustarem aos dados observados.

Uma característica importante das RNA é a capacidade de aprendizado automático a partir dos dados. Isso significa que, ao contrário de abordagens estatísticas tradicionais, as redes neurais podem aprender padrões complexos e representações abstratas diretamente dos dados, sem a necessidade de especificar explicitamente a forma da função a ser ajustada (Zador, 2019).

Neste trabalho, é proposta uma abordagem comparativa para a estimativa futura de variáveis hidrológicas utilizando técnicas de aprendizagem de máquina. O objetivo é analisar e comparar diferentes modelos de aprendizado de máquina aplicados a conjuntos de dados hidrológicos, buscando identificar qual técnica ou combinação de técnicas apresenta o melhor desempenho na previsão dessas variáveis. As abordagens foram selecionadas com base em sua capacidade de lidar com a complexidade e não linearidade dos processos hidrológicos, conforme discutido em estudos anteriores.

Essa pesquisa é relevante e justificada pela necessidade de avançar o conhecimento na área de estimativa hidrológica, buscando aprimorar as técnicas utilizadas atualmente e fornecer informações mais precisas para a gestão dos recursos hídricos. Ao obter previsões mais confiáveis de precipitação, temperatura, umidade do ar, velocidade do vento e outras variáveis hidrológicas é possível melhorar a tomada de decisões em relação ao uso sustentável da água e mitigação de desastres naturais, como enchentes. O impacto potencial deste trabalho abrange desde a redução de riscos e danos causados por eventos hidrológicos extremos até a otimização do uso dos recursos hídricos, promovendo a sustentabilidade e a resiliência dos sistemas hidrológicos.

---

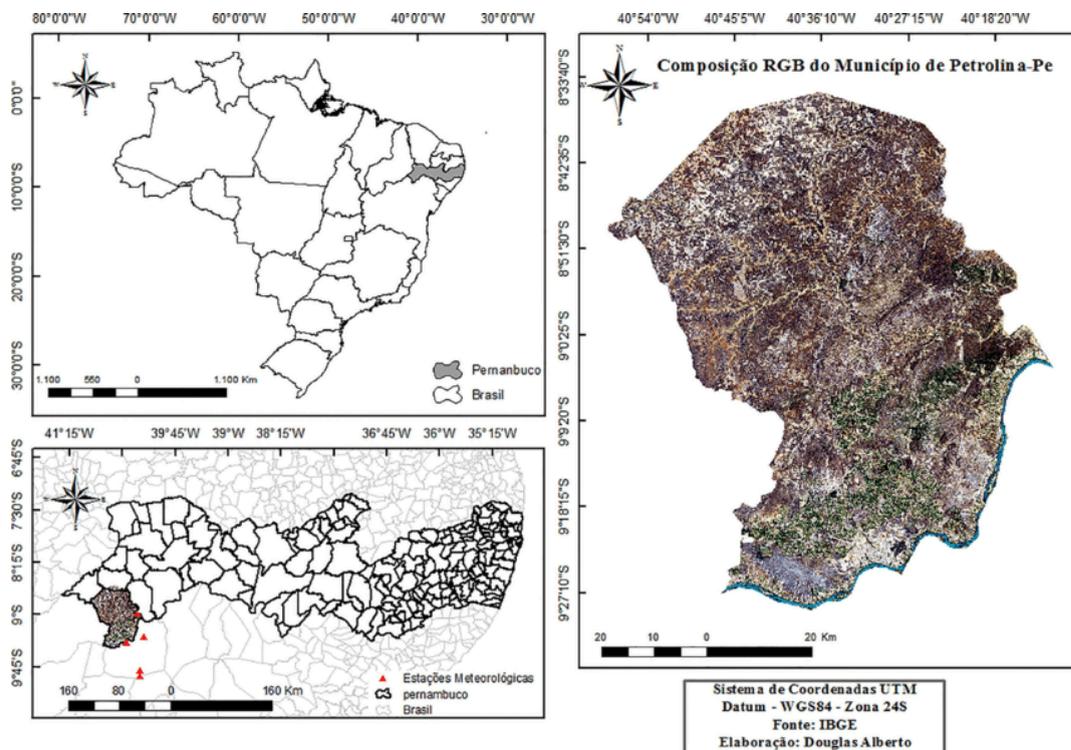
## **2 METODOLOGIA**

### **2.1 Área de estudo**

Foram selecionados os dados de Petrolina, município pernambucano que se destaca tanto social quanto economicamente. O município abrange uma área de 4561,872 km<sup>2</sup> e possui uma agricultura bastante desenvolvida, beneficiada pela sua localização às margens do Rio São Francisco. Devido às baixas precipitações, temperaturas elevadas, invernos curtos e à vegetação predominante de fragmentos do domínio fitogeográfico Caatinga, Petrolina desperta interesse científico e econômico para o estudo de sua dinâmica e comportamento temporal, o que justifica a escolha desse município.

Localizada no Vale do Rio São Francisco, no Nordeste do Brasil, Petrolina é caracterizada por um clima do tipo BSw<sup>h</sup>, conforme a classificação de Köppen, ou seja, refere-se a um clima árido com verões secos, invernos secos e temperaturas acima de 20°C durante o mês mais quente, e precipitações abaixo de 30mm durante o mês mais quente do ano. Essa região enfrenta uma demanda hídrica significativa devido à expansão da agricultura irrigada e ao rápido crescimento populacional observado nas últimas décadas.

**Figura 1 - Localização do município de Petrolina**



Fonte: Silva *et al.* (2018)

O posto de coleta de dados em Petrolina, devidamente identificado como A307-INMET, está estrategicamente posicionado geograficamente (Latitude: -9.388, Longitude: -40.523, Altitude: 372.72). Os registros foram iniciados em fevereiro de 2003 e se estendem até janeiro de 2023, com medições horárias que abrangem parâmetros como a data da medição, horário da medição, precipitação total para o intervalo temporal, pressão atmosférica ao nível da estação, pressão atmosférica reduzida ao nível do mar, pressão atmosférica máxima no intervalo temporal, pressão atmosférica mínima no intervalo temporal, radiação global, temperatura do ar no bulbo seco, temperatura do ar no ponto de orvalho, temperatura máxima

horária, temperatura mínima, temperatura no ponto de orvalho, temperatura mínima no ponto de orvalho, umidade relativa do ar, umidade relativa máxima, umidade relativa mínima, direção do vento, velocidade média do vento, velocidade máxima, dentre outros parâmetros.

A área de pesquisa abrange o polo de desenvolvimento em Petrolina, PE, onde as características climáticas, incluindo uma exposição intensa à radiação solar, temperaturas elevadas e uma estação chuvosa concentrada entre fevereiro e maio, exercem uma influência direta sobre a evapotranspiração e a disponibilidade hídrica (Barbosa et al., 2023). Com mais da metade da extensão territorial da bacia do Rio São Francisco localizada na região semiárida do Nordeste, a compreensão desses fatores é fundamental para a gestão eficaz dos recursos hídricos (Da Silva et al., 2023). O desfecho das correntes de ar na região resulta em um amplo corredor, mais seco no sentido NE-SW, onde se encontra o Vale do Rio São Francisco.

### 2.3 Tratamento dos dados

Inicialmente, adotou-se a estratégia de preenchimento de registros temporais que continham dados inconsistentes com a realidade, substituindo pelo valor médio correspondente ao respectivo parâmetro. Essa abordagem foi aplicada visando mitigar lacunas nos dados. Entretanto, é importante destacar que essa ação induziu uma ligeira volatilidade nos resultados, com uma magnitude da ordem de  $10^{-7}$ . Vale ressaltar que essa volatilidade é decorrente do próprio processo de preenchimento e deve ser interpretada considerando essa escala.

Registros nos quais não era possível ou que não faziam sentido preencher com o valor médio geral da coluna foram eliminados de maneira integral da tabela. Essa medida foi adotada para assegurar a integridade e a coerência dos dados analisados. Após a execução dos procedimentos mencionados anteriormente, restaram um total de 122.444 registros temporais no banco de dados.

Quanto ao processo de ajuste das informações coletadas, foi empregada a Equação 1, que descreve a técnica de normalização do logaritmo dos dados:

$$E'_i = \frac{\ln(E_i+1) - \text{média}(E \ln)}{S \ln} \quad (1)$$

Onde:

$E'_i$  a informação temporal normalizada para o tempo  $i$ ;

$E_i$  é a informação temporal no tempo  $i$ ;

$média(Eln)$  a média do  $\ln(E_i+1)$  de treino;  
 $Sln$  o desvio padrão do  $\ln(E_i)$  de treino.

Esse método de normalização do logaritmo busca padronizar os dados, levando em consideração a média e o desvio padrão do logaritmo natural das observações de treinamento. Ao aplicar essa técnica, os valores são ajustados de forma a considerar a distribuição estatística subjacente dos dados, contribuindo para uma análise mais coerente e comparável entre diferentes registros temporais. A abordagem de normalização do logaritmo é especialmente útil para reduzir distorções decorrentes de variações extremas e, assim, promover resultados mais robustos e confiáveis em análises subsequentes.

No tocante aos dados de entrada e saída dos modelos, procedeu-se ao cálculo e organização das correlações possíveis entre as variáveis, empregando o método de Pearson conforme delineado por Cohen et al. (2009). Esse método é um referencial estabelecido para avaliar a relação entre variáveis, oferecendo insights sobre possíveis associações lineares. Com o intuito de calcular essas correlações, foi utilizada a função integrada da biblioteca *Pandas* do Python.

Os coeficientes de correlação resultantes foram dispostos em três matrizes (apêndice A, B e C), oferecendo uma visualização clara das relações entre as variáveis coletadas. Valores próximos de 1 indicam uma correlação positiva forte, sugerindo uma relação direta entre as variáveis. Por outro lado, valores próximos de -1, indicam correlações mais fracas e, em alguns casos, negativas, onde uma variável tende a aumentar enquanto a outra diminui.

Para investigar as relações não lineares entre variáveis, recorreu-se ao coeficiente de correlação não linear de Spearman. Esse coeficiente é empregado para avaliar a associação entre duas variáveis, concentrando-se não na linearidade, mas sim na ordem das classificações dessas variáveis (Yu; Hutson, 2024). Em outras palavras, o coeficiente de correlação de Spearman mensura a relação monotônica entre duas variáveis, determinando se, à medida que uma variável aumenta, a outra também aumenta, independentemente de uma relação linear direta. Essa abordagem revela-se particularmente útil quando a relação entre as variáveis não se adequa estritamente a um padrão linear, podendo manifestar-se em formas curvas ou não lineares.

Segundo Yu e Hutson (2024), os valores obtidos variam de -1 a 1. Um valor próximo de 1 indica uma correlação forte e positiva, indicando que, conforme uma

variável aumenta, a outra também tende a aumentar, seguindo uma tendência monotônica. Por outro lado, um valor próximo de -1 sugere uma correlação forte e negativa, denotando que, à medida que uma variável aumenta, a outra tende a diminuir, seguindo uma tendência decrescente monotônica. Um valor próximo de 0 sugere uma correlação fraca ou nula, indicando uma ausência de relação monotônica entre as variáveis.

## 2.4 Elaboração dos algoritmos utilizados

No âmbito do desenvolvimento dos algoritmos de aprendizado de máquina, foram adotados os recursos disponibilizados pela biblioteca Scikit-Learn. De acordo com Rajamani e Iyer (2023), esta biblioteca é uma ferramenta de código aberto na linguagem de programação Python, projetada para a implementação de algoritmos e métricas voltados a projetos que abrangem aprendizagem de máquina.

Para este estudo, uma série de algoritmos relevantes foram selecionados para análise. Entre estes, destacam-se a aplicação da *artificial neural networks* (ANN), *support vector machines* (SVM), *decision trees* (DT), *random forests* (RF), e *gradient boosting* (GB). Essa seleção contempla uma variedade de abordagens que abrangem desde métodos mais tradicionais, como árvores de decisão e máquinas de suporte vetorial, até técnicas mais avançadas, como as redes neurais artificiais.

A escolha desses algoritmos se baseou em sua capacidade de lidar com a complexidade e a não linearidade dos dados hidrológicos, bem como em sua eficácia na realização de tarefas de previsão e modelagem. A biblioteca Scikit-Learn proporciona uma interface amigável e acessível para a implementação desses algoritmos, permitindo que a análise seja conduzida com rigor científico e praticidade.

### 2.4.1 Máquina de suporte vetorial

Para realizar a regressão utilizando o algoritmo de Máquinas de Suporte Vetorial (SVM), temos:

- Importação de bibliotecas: Primeiramente, foi importado as bibliotecas necessárias para realizar as etapas subsequentes, incluindo o *numpy*, *pandas*, *scipy.stats*, *sklearn* e funções personalizadas para cálculo de métricas.
- Preparação dos dados: Os conjuntos de dados de treino e teste foram devidamente preparados. Foram importados os dados de entrada ( $x_{\text{treino}}$  e  $x_{\text{teste}}$ ) e os respectivos rótulos alvo ( $y_{\text{treino}}$  e  $y_{\text{teste}}$ ).

- Configuração dos parâmetros do algoritmo: Foi definido os parâmetros para o algoritmo de Máquinas de Suporte Vetorial (SVM) por meio de um dicionário `parametros_svm`. Os parâmetros incluem o tipo de *kernel* (*linear*, *poly*, *rbf*, *sigmoid*), grau do polinômio (*degree*), parâmetros relacionados à função de *kernel* (*gamma*, *tol*, *epsilon*, *shrinking*, *verbose*, *max\_iter*), entre outros.
- Instanciação do modelo e *randomized search*: Foi criada uma instância do modelo de regressão de Máquinas de Suporte Vetorial (`modelo_svm = SVR()`) e, em seguida, aplicado a busca aleatória (`RandomizedSearchCV`) para encontrar os melhores parâmetros do modelo. O processo de busca foi guiado pelas configurações definidas anteriormente, como o número de iterações (`n_iter`) e a pontuação a ser otimizada (*scoring*).
- Treinamento e avaliação: O modelo foi treinado utilizando os dados de treino (`x_treino` e `y_treino`) por meio do método *fit()*. Em seguida, obtivemos os resultados da busca aleatória e identificamos os melhores parâmetros (`indice_melhores_parametros`).

#### 2.4.2 Árvore de decisão

A abordagem metodológica do algoritmo de Árvore de Decisão segue um processo sequencial detalhado. Em primeiro lugar, realizou-se a importação das bibliotecas essenciais para as etapas subsequentes do processo. Foram incorporadas as bibliotecas *numpy*, *pandas*, *scipy.stats*, *sklearn* e funções personalizadas destinadas ao cálculo das métricas de avaliação.

A preparação dos dados desempenhou um papel crucial na metodologia. Os conjuntos de dados de treinamento e teste foram adequadamente preparados, com a importação dos dados de entrada (`x_treino` e `x_teste`) e seus respectivos rótulos alvo (`y_treino` e `y_teste`).

Em seguida, foram configurados os parâmetros do algoritmo de Árvore de Decisão por meio de um dicionário denominado "parametros\_arvore". Este dicionário engloba uma série de parâmetros cruciais para o desempenho do modelo, tais como o critério de divisão, a profundidade máxima da árvore, o número mínimo de amostras para divisão e folhas, e outras configurações.

A instanciação do modelo foi realizada através da criação de um objeto de regressão da Árvore de Decisão. Em seguida, aplicou-se a busca aleatória (`RandomizedSearchCV`) para encontrar os melhores parâmetros do modelo. A

busca foi guiada pelas configurações predefinidas, como o número de iterações, o escore a ser otimizado e a validação cruzada.

Após a busca aleatória, o modelo foi treinado utilizando os dados de treinamento, por meio do método *fit()*. Em seguida, foram obtidos os resultados da busca aleatória, identificando os melhores parâmetros por meio do índice correspondente.

### 2.4.3 Floresta aleatória

O procedimento adotado no algoritmo de Floresta Aleatória percorre etapas sequenciais, abrangendo desde a configuração dos parâmetros do algoritmo até o treinamento do modelo:

- Importação de bibliotecas: Inicialmente, foram realizadas as importações das bibliotecas cruciais para a execução das fases subsequentes. Isso abarca a incorporação de módulos como *scipy.stats*, *sklearn*, além de funções personalizadas para o cálculo de métricas.
- Preparação dos dados: Os conjuntos de treinamento e teste foram meticulosamente preparados. Nessa etapa, os dados de entrada (*x\_treino* e *x\_teste*) e os rótulos alvo correspondentes (*y\_treino* e *y\_teste*) foram importados e organizados.
- Configuração dos parâmetros do algoritmo: Definiu-se, por meio do dicionário "parametros\_floresta", os parâmetros pertinentes ao algoritmo de Floresta Aleatória. Dentre esses parâmetros, destacam-se o número de estimadores na floresta (*n\_estimators*), critérios de divisão, a profundidade máxima da árvore (*max\_depth*), entre outros.
- Instanciação do modelo e busca aleatória: Realizou-se a instanciação do modelo de regressão de Floresta Aleatória (*modelo\_floresta = RandomForestRegressor()*). Na sequência, empregou-se a técnica de busca aleatória (*RandomizedSearchCV*) para identificação dos parâmetros ótimos do modelo. Tal processo de busca incorporou parâmetros como o número de iterações (*n\_iter*) e a pontuação alvo para otimização (*scoring*).
- Treinamento e avaliação: O modelo foi submetido ao treinamento utilizando os conjuntos de treinamento (*x\_treino* e *y\_treino*), por intermédio do método *fit()*. Em seguida, procedeu-se à análise dos resultados da busca aleatória, culminando na identificação dos parâmetros de maior relevância.

### 2.4.4 Reforço por gradiente

A metodologia adotada para a aplicação do algoritmo de *Gradient Boosting* no contexto da regressão foi conduzida seguindo uma abordagem estruturada e sistemática. O processo compreendeu diversas etapas, desde a preparação dos dados até a avaliação das métricas de desempenho do modelo resultante. A seguir, descreve-se detalhadamente a metodologia empregada:

- Importação de bibliotecas: Inicialmente, foram importadas as bibliotecas essenciais para as análises subsequentes. Isso incluiu a biblioteca *scipy.stats* para funcionalidades estatísticas e o módulo *sklearn* para a implementação do algoritmo de *Gradient Boosting*. Além disso, foram incorporadas funções personalizadas para o cálculo de métricas relevantes.
- Preparação dos dados: Os conjuntos de dados de treinamento e teste foram cuidadosamente preparados. Os dados de entrada (*x\_treino* e *x\_teste*) foram importados juntamente com os rótulos alvo correspondentes (*y\_treino* e *y\_teste*).
- Configuração dos parâmetros do algoritmo: Os parâmetros do algoritmo de *Gradient Boosting* foram definidos por meio de um dicionário denominado "parametros\_gradiente". Esse conjunto de parâmetros abrange aspectos como a função de perda (*loss*), a taxa de aprendizado (*learning\_rate*), o número de estimadores (*n\_estimators*), entre outros.
- Instanciação do modelo e busca aleatória: Uma instância do modelo de regressão por *Gradient Boosting* foi criada (*modelo\_gradiente = GradientBoostingRegressor()*). Em seguida, empregou-se a técnica de busca aleatória (*RandomizedSearchCV*) para determinar os melhores parâmetros do modelo. A busca aleatória considerou fatores como o número de iterações (*n\_iter*) e a pontuação a ser otimizada (*scoring*).
- Treinamento e avaliação: O modelo de Gradient Boosting foi treinado utilizando os conjuntos de treinamento (*x\_treino* e *y\_treino*) por meio do método *fit()*. Os resultados da busca aleatória foram analisados para identificar os parâmetros de maior relevância para o modelo.

#### **2.4.5 Redes neurais artificiais**

A aplicação de RNA no contexto da regressão foi conduzida por meio de uma metodologia estruturada e sistemática. A seguir, descreve-se detalhadamente o procedimento adotado para a implementação e avaliação das RNA:

- Definição do espaço de parâmetros: Inicialmente, foi estabelecido um espaço de parâmetros abrangente que engloba diferentes aspectos das RNAs, tais como o tamanho das camadas ocultas (*hidden\_layer\_sizes*), a função de ativação (*activation*), o otimizador (*solver*), entre outros. O espaço de parâmetros foi elaborado com base em referências especializadas e considerou múltiplas opções para cada componente das RNAs.
- Instanciação do modelo e randomized search: Foi criada uma instância do modelo de Regressão com Redes Neurais Artificiais (*modelo\_neural\_regressor = MLPRegressor()*) para ser utilizada na aplicação do algoritmo. Em seguida, a técnica de busca aleatória (*RandomizedSearchCV*) foi empregada para explorar e encontrar a combinação mais adequada de parâmetros dentro do espaço previamente definido.
- Treinamento e avaliação: A RNA foi treinada utilizando o conjunto de treinamento (*x\_treino* e *y\_treino*) por meio do método *fit()*. Durante o processo de treinamento, foram avaliadas diferentes combinações de parâmetros em busca daquela que proporcionasse o melhor desempenho de acordo com a métrica de erro quadrático médio negativo (*neg\_mean\_squared\_error*).

## 2.5 Otimização dos algoritmos

A abordagem metodológica empregada consistiu na divisão dos dados em duas partes aleatórias: 70% dos dados foram destinados ao treinamento dos modelos, enquanto os restantes 30% foram reservados para o teste e validação dos modelos. As configurações adotadas nas funções e algoritmos de aprendizagem de máquina foram cuidadosamente selecionadas e mantidas consistentes para todas as variáveis consideradas no processo de previsão. Essa abordagem foi adotada após uma série de experimentos e ajustes, visando tanto a otimização dos resultados quanto a padronização das técnicas empregadas para atingir os objetivos do estudo.

O algoritmo de otimização utilizado neste estudo é o *RandomizedSearchCV*, disponível na biblioteca *Scikit-Learn* (Rajamani; Iyer, 2023). O *RandomizedSearchCV* é uma ferramenta poderosa que realiza uma busca aleatória nos hiperparâmetros de um estimador de aprendizado de máquina, com o objetivo

de encontrar a melhor combinação de parâmetros que otimize o desempenho do modelo (Arauna, 2023).

O `RandomizedSearchCV` implementa os métodos "fit" e "score", permitindo a realização de ajustes otimizados de hiperparâmetros e a avaliação do desempenho do modelo em um conjunto de dados de teste. Além disso, ele também implementa outros métodos, como "predict", "predict\_proba", "decision\_function", "transform" e "inverse\_transform", desde que esses métodos sejam implementados no estimador usado.

Diferentemente do `GridSearchCV`, que explora todas as combinações possíveis de hiperparâmetros, o `RandomizedSearchCV` realiza uma amostragem aleatória de um número fixo de configurações de parâmetros a partir das distribuições especificadas. Esse número é definido pelo parâmetro "n\_iter" (Vishnu *et al.*, 2023).

Os hiperparâmetros a serem otimizados podem ser fornecidos como um dicionário ou lista de dicionários, onde as chaves são os nomes dos parâmetros e os valores são distribuições ou listas de valores a serem testados. O algoritmo realiza a amostragem de acordo com essas distribuições, buscando encontrar os valores que maximizam o desempenho do modelo.

Segundo Azmain *et al.* (2022), para avaliar o desempenho do modelo, o `RandomizedSearchCV` permite a escolha de uma estratégia de pontuação, que pode ser uma única métrica ou uma combinação de várias métricas. O resultado é uma avaliação do modelo para cada combinação de hiperparâmetros testada.

Os resultados do processo de otimização, incluindo os melhores parâmetros encontrados, a pontuação correspondente e outras informações relevantes, são armazenados na estrutura de dados "cv\_results\_", que pode ser convertida em um `DataFrame` do `Pandas` para análise mais detalhada.

## 2.6 Métricas

O coeficiente de eficiência de Nash-Sutcliffe (NSE) (Equação 2), é calculado como a proporção do erro quadrático médio entre o valor observado e simulado em relação à variância do valor observado. O NSE varia entre -infinito e 1, sendo que um valor de 1 indica um acordo perfeito entre o escoamento observado e simulado. Esse coeficiente é empregado como uma métrica para avaliar a concordância e qualidade das previsões entre os valores observados e os valores simulados.

$$NSE(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1} (y_i - \text{mean}(y))^2} \quad (2)$$

Onde:

$y_i$  é o valor medido,  $\hat{y}_i$  é o valor estimado,  $\text{mean}(y)$  é a média dos valores medidos,  $N$  é o número de registros medidos.

O Índice de Willmott (WI) (Equação 3) é uma medida estatística empregada para avaliar o desempenho de um modelo de previsão, particularmente no contexto de variáveis hidrológicas ou relacionadas ao clima. O WI compara a precisão de um modelo com a precisão de um modelo de referência que simplesmente prevê o valor médio da variável observada. O WI varia entre 0 e 1, com um valor de 1 indicando acordo perfeito entre os valores previstos e observados. Um valor de 0 indica que os valores previstos não são melhores do que a previsão da média dos valores observados. Este é frequentemente utilizado em campos relacionados à hidrologia e ao clima para avaliar a precisão de modelos que preveem variáveis como precipitação, temperatura e evapotranspiração. Trata-se de uma ferramenta útil para comparar o desempenho de diferentes modelos ou diferentes métodos de estimativa de uma variável (Silva *et al.*, 2017).

$$WI(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1} (|\hat{y}_i - \text{mean}(y)| + |y_i - \text{mean}(y)|)^2} \quad (3)$$

Onde:

$y_i$  é o valor medido,  $\hat{y}_i$  é o valor estimado,  $\text{mean}(y)$  é a média dos valores medidos,  $N$  é o número de registros medidos.

O Índice de Correlação de Pearson (R) (Equação 4), também conhecido como Coeficiente de Correlação de Pearson, é uma medida estatística que quantifica a intensidade e a direção da relação linear entre duas variáveis. É representado pelo símbolo "R" e varia entre -1 e +1 (Thieu *et al.*, 2023).

Um valor de +1 indica uma relação linear positiva perfeita entre as duas variáveis, enquanto um valor de -1 indica uma relação linear negativa perfeita. Um valor de 0 indica a ausência de relação linear entre as duas variáveis. O coeficiente de correlação de Pearson pode ser utilizado para determinar a intensidade e a direção da relação entre duas variáveis.

O coeficiente de correlação de Pearson é comumente empregado em diversos campos, incluindo ciências sociais, economia e engenharia, para estudar a relação entre duas variáveis. É importante ressaltar que o coeficiente de correlação de Pearson mede apenas relações lineares entre variáveis, podendo não capturar outros tipos de relações, como relações não-lineares ou não-monotônicas.

$$R(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} ((y_i - \text{mean}(y)) \times (\hat{y}_i - \text{mean}(\hat{y}_i)))}{\sqrt{\sum_{i=0}^{N-1} (y_i - \text{mean}(y))^2} \times \sqrt{\sum_{i=0}^{N-1} (\hat{y}_i - \text{mean}(\hat{y}_i))^2}} \quad (4)$$

Onde:

$y_i$  é o valor medido,  $\hat{y}_i$  é o valor estimado,  $\text{mean}(y)$  é a média dos valores medidos,  $\text{mean}(\hat{y}_i)$  é a média dos valores estimado,  $N$  é o número de registros medidos.

Segundo Moraes *et al.* (2023), com o índice de desempenho do modelo (c) (Equação 5), é possível testar a precisão e a exatidão dos resultados gerados, avaliados conforme a tabela 1.

**Tabela 1** - Critério de interpretação do desempenho pelo índice "c"

| Valor de "c" | Desempenho |
|--------------|------------|
| > 0,85       | Ótimo      |
| 0,76 a 0,85  | Muito Bom  |
| 0,66 a 0,75  | Bom        |
| 0,61 a 0,65  | Mediano    |
| 0,51 a 0,60  | Sofrível   |
| 0,41 a 0,50  | Mau        |
| ≤ 0,40       | Péssimo    |

Fonte: Moraes *et al.* (2023)

$$c = WI \times R \quad (5)$$

Onde:

WI é o índice de Willmott, R é o índice de correlação de Pearson.

O Índice de Eficiência Kling-Gupta (KGE) (Equações 6 a 8) é uma medida estatística utilizada para avaliar o desempenho de modelos hidrológicos. Este foi proposto para superar as limitações de outras métricas, como a Eficiência Nash-Sutcliffe e o R-quadrado, que se concentram apenas na reprodução da média e variância dos dados observados (Thieu *et al.*, 2023).

O KGE combina três métricas estatísticas: coeficiente de correlação, razão de variabilidade e razão de viés, em uma única medida de desempenho do modelo. A faixa de variação do KGE é  $(-\infty, 1]$ , onde um valor de 1 indica concordância

perfeita entre as previsões do modelo e os dados observados. Este avalia não apenas a precisão das previsões do modelo, mas também sua capacidade de reproduzir a variabilidade e o cronograma dos dados observados. Este tem sido amplamente utilizado em hidrologia e engenharia de recursos hídricos para avaliar o desempenho de modelos hidrológicos na simulação de vazão de riachos, recarga de aquíferos e parâmetros de qualidade da água (Thieu *et al.*, 2023).

$$KGE(y, \hat{y}) = 1 - \sqrt{(R(y, \hat{y}) - 1)^2 + (\beta(y, \hat{y}) - 1)^2 + (\gamma(y, \hat{y}) - 1)^2} \quad (6)$$

$$\beta = \text{Razão de Viés} = \frac{\mu_{\hat{y}}}{\mu_y} \quad (7)$$

$$\gamma = \text{Relação de Variabilidade} = \frac{CV_{\hat{y}}}{CV_y} = \frac{\frac{\sigma_{\hat{y}}}{\mu_{\hat{y}}}}{\frac{\sigma_y}{\mu_y}} \quad (8)$$

Onde:

$CV$  é o coeficiente de variação dos valores estimados ( $CV_{\hat{y}}$ ) e medidos ( $CV_y$ ),  $\sigma$  é o desvio padrão dos valores estimados ( $\sigma_{\hat{y}}$ ) e medidos ( $\sigma_y$ ),  $\mu$  é a média dos valores estimados ( $\mu_{\hat{y}}$ ) e medidos ( $\mu_y$ ),  $R$  é o índice de correção de Pearson.

O Coeficiente de Correlação de Pearson ( $R^2$ s) (Equação 9) ao Quadrado, também conhecido como  $R$  quadrado ( $R^2$ ) ou  $R^2$ s, é uma medida estatística que quantifica a relação entre duas variáveis ao quadrado. A pontuação ideal é 1, sendo que um valor maior indica um melhor desempenho. A faixa de variação é [0, 1].

A maioria dos tutoriais online e até mesmo a biblioteca scikit-learn equivocadamente utilizam a notação errada para  $R^2$ s e  $R^2$ .  $R^2 = R^2$ s =  $R$  quadrado, o que induz ao erro de considerar que se trata do Coeficiente de Correlação de Pearson elevado ao quadrado. Entretanto,  $R^2$  é o Coeficiente de Determinação, conforme explicado em Negelkerke *et al.* (1991).

$$R^2s(y, \hat{y}) = \left[ \frac{\sum_{i=0}^{N-1} ((y_i - \text{mean}(y)) \times (\hat{y}_i - \text{mean}(\hat{y})))}{\sqrt{\sum_{i=0}^{N-1} (y_i - \text{mean}(y))^2} \times \sqrt{\sum_{i=0}^{N-1} (\hat{y}_i - \text{mean}(\hat{y}))^2}} \right]^2 \quad (9)$$

Onde:

$y_i$  é o valor medido,  $\hat{y}_i$  é o valor estimado,  $mean(y)$  é a média dos valores medidos,  $mean(\hat{y}_i)$  é a média dos valores estimado,  $N$  é o número de registros medidos.

Segundo Nguyen *et al.* (2019), o Erro Quadrático Médio da Raiz (RMSE) (Equação 10) é uma medida estatística frequentemente utilizada para avaliar a precisão de um modelo de previsão, como um modelo de regressão ou um modelo de série temporal. Ele mede a diferença entre os valores previstos e os valores reais nas unidades da variável de resposta. O cálculo do RMSE é feito como a raiz quadrada da média dos quadrados das diferenças entre os valores previstos e os valores reais. Uma pontuação mais próxima de 0 é melhor. A faixa de variação é  $[0, +\infty)$ .

O RMSE é uma medida amplamente utilizada para avaliar a precisão das previsões, pois é sensível tanto à magnitude quanto à direção dos erros. Um RMSE mais baixo indica maior precisão na previsão. No entanto, ele apresenta uma desvantagem por não ser normalizado, ou seja, depende da escala da variável de resposta. Portanto, é difícil comparar os valores de RMSE entre diferentes conjuntos de dados com escalas distintas (Nguyen *et al.*, 2019). O RMSE é comumente empregado em diversas áreas, incluindo finanças, economia e engenharia, para avaliar o desempenho de modelos de previsão (Nguyen *et al.*, 2019).

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{N}} \quad (10)$$

Onde:

$y_i$  é o valor medido,  $\hat{y}_i$  é o valor estimado,  $N$  é o número de registros medidos.

Para avaliar de maneira abrangente as métricas calculadas para um algoritmo como um todo, foi introduzida uma expressão desenvolvida no âmbito deste estudo, denominada de "Somatório das Métricas" (Equação 13). Essa expressão é definida pela Equação 11:

$$SM = \sum_{i=1}^M |Métrica_i - Situação\ perfeita\ da\ métrica_i| \quad (11)$$

Onde, as métricas calculadas para o algoritmo em análise são subtraídas de seus valores ideais, representando uma situação de desempenho perfeito. O somatório dos valores absolutos resultantes das diferenças entre todas as métricas adotadas

busca minimizar-se, aproximando-se o máximo possível de zero. Com relação ao conjunto de equações avaliativas menores do que  $N=2$ , a expressão torna-se obsoleta.

---

### 3 RESULTADOS E DISCUSSÃO

A partir das matrizes geradas pelos métodos de Pearson (Apêndice A), Kendall (Apêndice B) e Spearman (Apêndice C), foram identificadas algumas combinações preliminares para testes pelos algoritmos desenvolvidos. Estas combinações foram selecionadas com base no critério do módulo da correlação calculada, sendo consideradas aquelas com valores iguais ou superiores a 0,6. Todas as combinações de variáveis mais pertinentes foram documentadas no Apêndice D, com os resultados mais relevantes das combinações detalhadas a seguir.

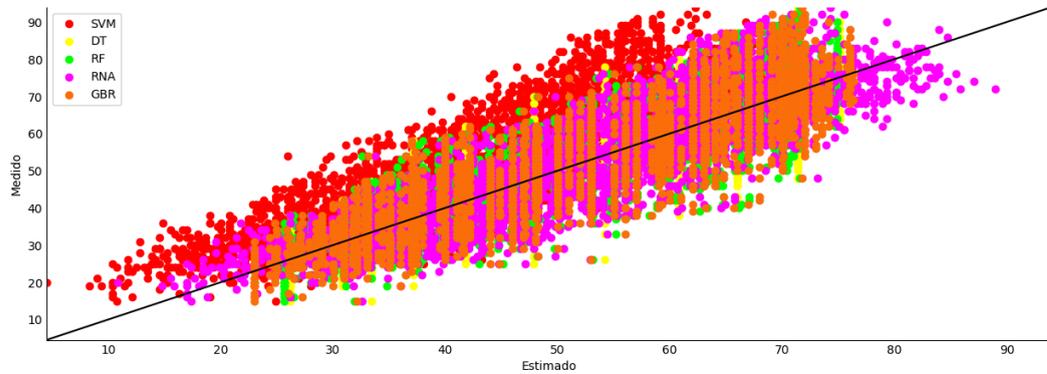
Considerado a temperatura do ar e a umidade relativa do ar como variável de entrada, focalizando no intervalo temporal de 0 dias para determinar o valor exato da variável de saída no momento do conhecimento da variável de entrada, observou-se a obtenção de resultados consistentes (Figuras 2 e 3) por meio das métricas avaliadas (Tabela 2). Para a maioria dos algoritmos processados, o NSE oscilou em torno de 0,75, enquanto o  $R^2$  se manteve próximo a 0,77. No entanto, o algoritmo SVM apresentou um desempenho inferior para essas configurações específicas de entrada e saída, revelando um NSE de 0,40 e um RMSE de 13,05, resultados distantes para o mesmo algoritmo avaliado por Kamir *et al.* (2019). Apesar disso, os demais parâmetros de avaliação demonstraram um desempenho bastante satisfatório e consistente entre os algoritmos considerados.

**Tabela 2** - Resultado das métricas para a combinação mencionada.

| <b>Modelo</b> | <b>NSE</b> | <b>d</b> | <b>r</b> | <b>c</b> | <b>KGE</b> | <b>R<sup>2</sup></b> | <b>RMSE</b> | <b>SM</b> |
|---------------|------------|----------|----------|----------|------------|----------------------|-------------|-----------|
| SVM           | 0.408      | 0.833    | 0.877    | 0.730    | 0.774      | 0.770                | 13.059      | 14.667    |
| DT            | 0.774      | 0.934    | 0.880    | 0.822    | 0.838      | 0.775                | 8.063       | 9.040     |
| RF            | 0.780      | 0.935    | 0.884    | 0.826    | 0.842      | 0.781                | 7.966       | 8.918     |
| GBR           | 0.775      | 0.934    | 0.881    | 0.823    | 0.840      | 0.777                | 8.048       | 9.019     |
| ANN           | 0.770      | 0.932    | 0.878    | 0.818    | 0.834      | 0.770                | 8.146       | 9.145     |

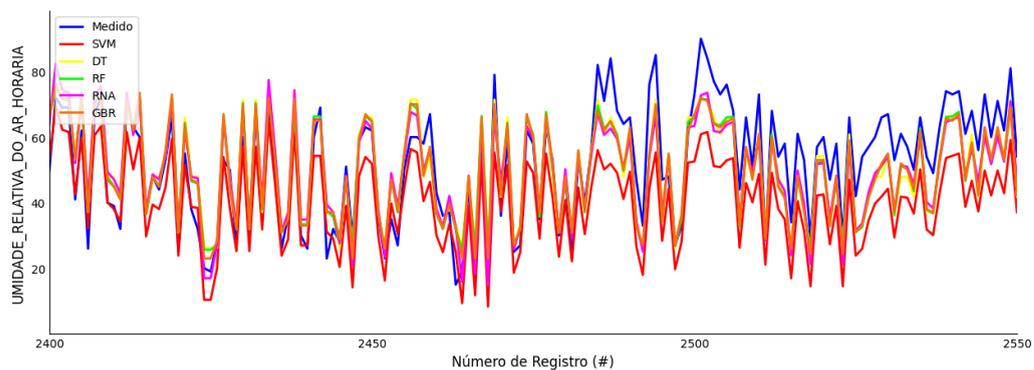
Fonte: Autor (2024)

**Figura 2** - Dispersão dos resultados previstos – valores de umidade do ar estimados a partir da temperatura do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 3** - Comparativo entre os valores reais (azul) e os estimados a partir da temperatura e da umidade do ar



Fonte: Do autor (2023)

Avaliando a relação entre a variável de entrada, temperatura do ar (bulbo seco) horária, e a variável de saída, também relacionada à temperatura do ar (bulbo seco) com um intervalo de três dias (Tabela 3). O algoritmo SVM obteve um NSE de 0,072, apresentou valores consideráveis nos coeficientes de correlação ( $d=0,699$ ,  $r=0,765$ ,  $c=0,535$ ,  $KGE=0,595$ ), parecidos com os resultados encontrados por Huang *et al.* (2019), mas com resultados menos favoráveis para o índice de determinação ( $R^2=0,585$ ) e RMSE (4,293). No entanto, os modelos DT, RF, GBR e ANN mostraram desempenho mais consistente (Figuras 4 e 5), com NSE variando entre 0,576 e 0,584, sendo o RF semelhante ao desempenho de Chen *et al.* (2020). Esses modelos apresentaram valores elevados nos coeficientes de correlação (superiores a 0,84) e índices de determinação  $R^2$  em torno de 0,58, além de RMSE abaixo de 3,

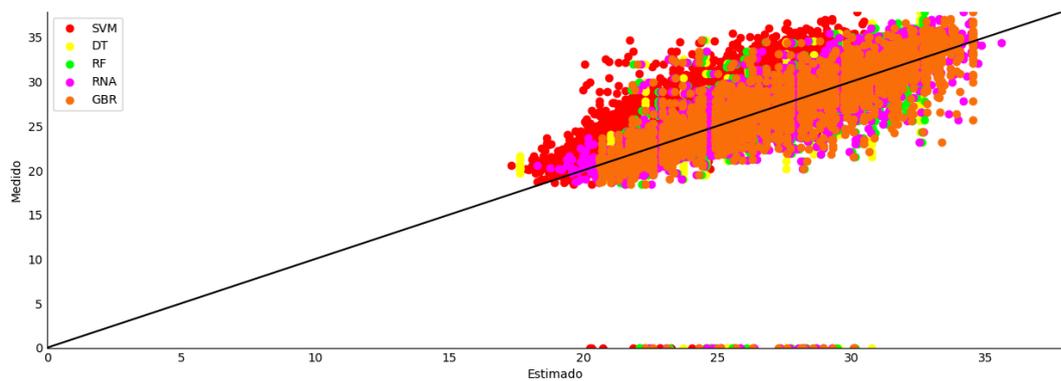
indicando resultados mais precisos e confiáveis.

**Tabela 3** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE   | d     | r     | c     | KGE   | R <sup>2</sup> | RMSE  | SM    |
|--------|-------|-------|-------|-------|-------|----------------|-------|-------|
| SVM    | 0.072 | 0.699 | 0.765 | 0.535 | 0.595 | 0.585          | 4.293 | 7.043 |
| DT     | 0.576 | 0.847 | 0.759 | 0.643 | 0.649 | 0.576          | 2.901 | 4.850 |
| RF     | 0.584 | 0.850 | 0.764 | 0.650 | 0.652 | 0.584          | 2.875 | 4.791 |
| GBR    | 0.583 | 0.860 | 0.765 | 0.658 | 0.689 | 0.586          | 2.878 | 4.738 |
| ANN    | 0.584 | 0.851 | 0.764 | 0.650 | 0.656 | 0.584          | 2.874 | 4.785 |

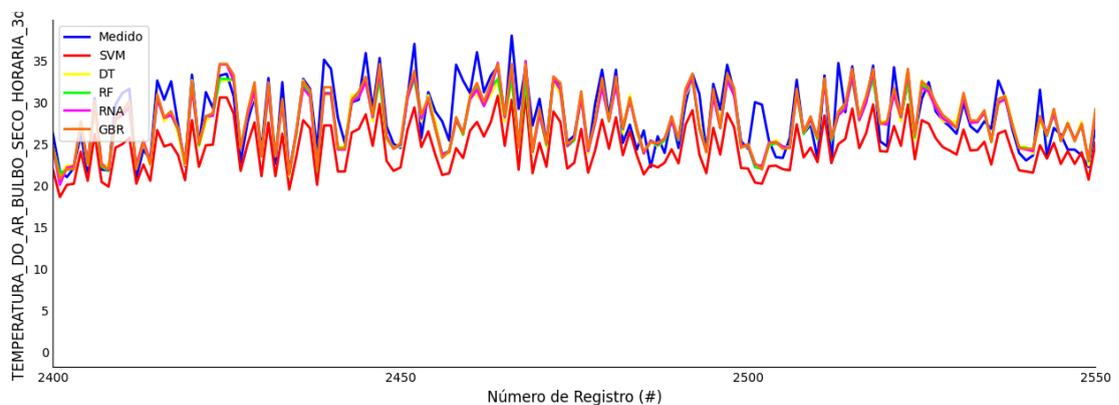
Fonte: Autor (2024)

**Figura 4** - Dispersão dos resultados previstos – valores da temperatura do ar estimados a partir da temperatura do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 5** - Comparativo entre os valores reais (azul) e os estimados a partir das temperaturas



Fonte: Do autor (2023)

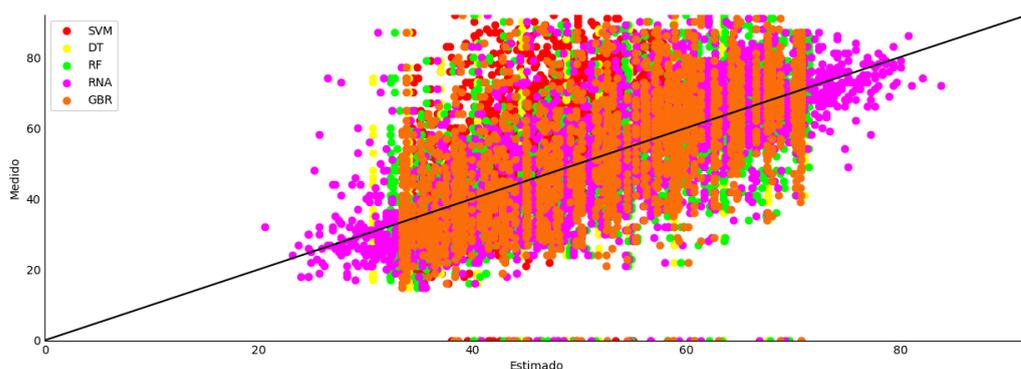
Para a relação entre a Temperatura do Ar (Bulbo Seco) como variável de entrada e a Umidade Relativa do Ar com um intervalo de 3 dias como variável de saída (figura 6 e 7) (Tabela 4). O modelo SVM registrou um NSE de 0,181, exibindo coeficientes de correlação ( $d=0,638$ ,  $r=0,705$ ,  $c=0,449$ ,  $KGE=0,358$ ) que indicam uma relação moderada entre as variáveis, com valores de  $d$  e  $r$  próximos aos obtidos por Dalla Torre *et al.* (2024). No entanto, seu desempenho, com um RMSE de 16,044, sugere uma precisão limitada na previsão da umidade relativa do ar. Em contraste, os modelos DT, RF, GBR e ANN apresentaram NSE mais elevados, variando de 0,499 a 0,506, com coeficientes de correlação acima de 0,71 e RMSE na faixa de 12,455 a 12,547, demonstrando previsões mais consistentes e com precisão considerável.

**Tabela 4** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE   | d     | r     | c     | KGE   | R <sup>2</sup> | RMSE   | SM     |
|--------|-------|-------|-------|-------|-------|----------------|--------|--------|
| SVM    | 0.181 | 0.638 | 0.705 | 0.449 | 0.358 | 0.497          | 16.044 | 19.202 |
| DT     | 0.506 | 0.815 | 0.712 | 0.580 | 0.590 | 0.507          | 12.455 | 14.745 |
| RF     | 0.506 | 0.815 | 0.712 | 0.580 | 0.588 | 0.506          | 12.455 | 14.748 |
| GBR    | 0.499 | 0.811 | 0.707 | 0.573 | 0.584 | 0.499          | 12.547 | 14.874 |
| ANN    | 0.502 | 0.811 | 0.709 | 0.575 | 0.584 | 0.503          | 12.506 | 14.822 |

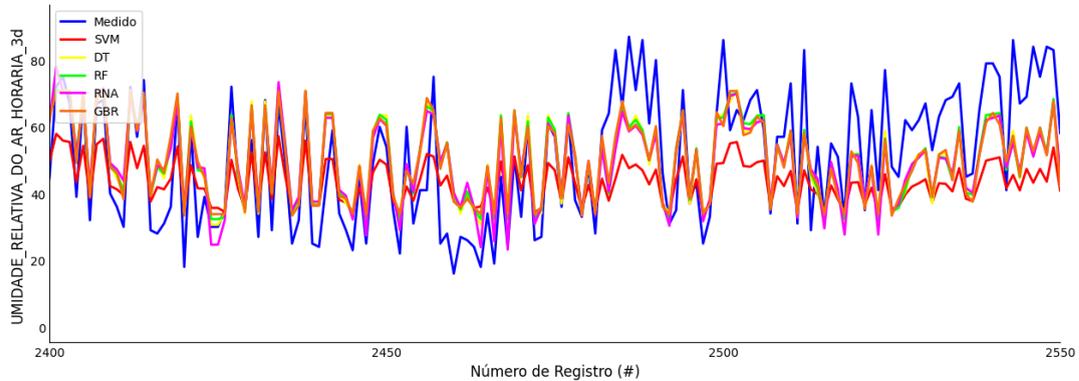
Fonte: Autor (2024)

**Figura 6** - Dispersão dos resultados previstos – valores da umidade do ar estimados a partir da temperatura do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 7** - Comparativo entre os valores reais (azul) e os estimados a partir da temperatura e da umidade do ar



Fonte: Do autor (2023)

Considerando a Temperatura do Ar (Bulbo Seco) como variável de entrada e a previsão da Temperatura do Ar (Bulbo Seco) com um intervalo de 7 dias como variável de saída, diferentes modelos foram aplicados para avaliação (figura 8 e 9). O modelo SVM exibiu um NSE negativo (-5,202), apontando para um desempenho substancialmente inferior em comparação com os demais modelos. Seus coeficientes de correlação ( $d=0,388$ ,  $r=0,627$ ,  $c=0,243$ ,  $KGE=0,437$ ) sugerem uma relação limitada entre as variáveis, enquanto o RMSE de 12,489 indica uma precisão insatisfatória na previsão. Por outro lado, os modelos DT, RF, GBR e ANN demonstraram NSE variando de 0,449 a 0,458, com coeficientes de correlação acima de 0,67 e RMSE na faixa de 3,692 a 3,723. Ao comparar o modelo ANN com com o mesmo algoritmo desenvolvido por Prieto *et al.* (2019), ambos para intervalos maiores de tempo, temos um desempenho muito semelhante.

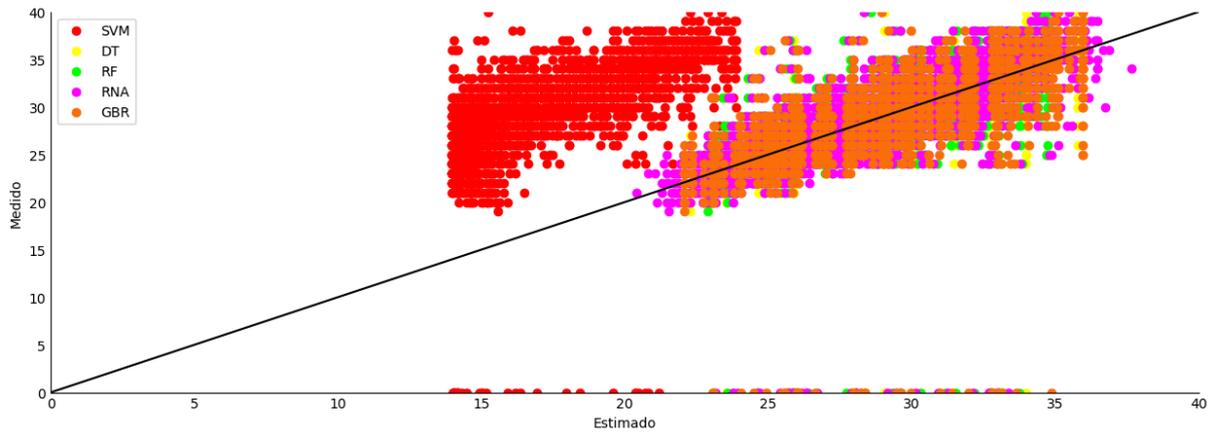
**Tabela 5** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE    | d     | r     | c     | KGE   | R <sup>2</sup> | RMSE   | SM     |
|--------|--------|-------|-------|-------|-------|----------------|--------|--------|
| SVM    | -5.202 | 0.388 | 0.627 | 0.243 | 0.437 | 0.393          | 12.489 | 21.601 |
| DT     | 0.453  | 0.783 | 0.673 | 0.528 | 0.545 | 0.453          | 3.710  | 6.275  |
| RF     | 0.456  | 0.784 | 0.676 | 0.530 | 0.544 | 0.457          | 3.698  | 6.250  |
| GBR    | 0.449  | 0.795 | 0.676 | 0.537 | 0.576 | 0.457          | 3.723  | 6.232  |

|     |       |       |       |       |       |       |       |       |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|
| ANN | 0.458 | 0.785 | 0.677 | 0.532 | 0.546 | 0.459 | 3.692 | 6.234 |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|

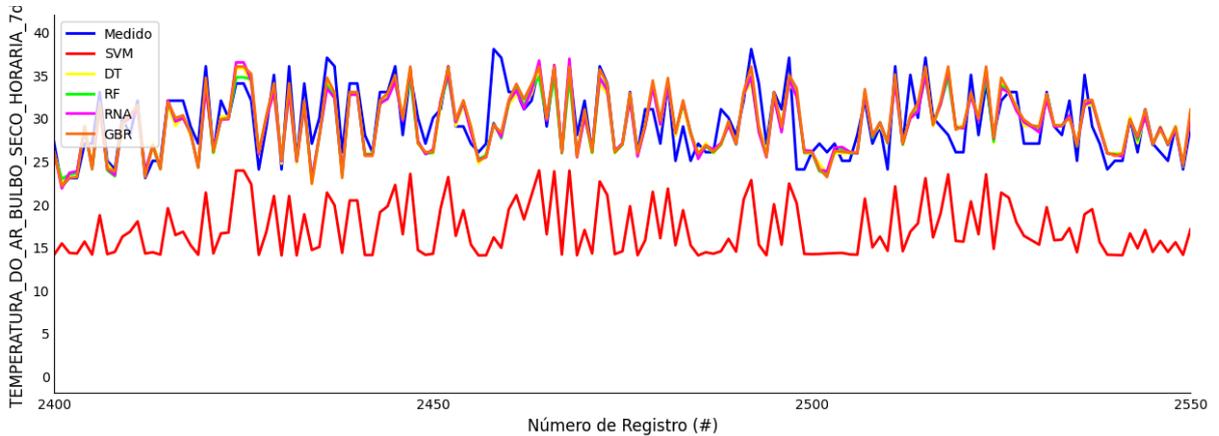
Fonte: Autor (2024)

**Figura 8** - Dispersão dos resultados previstos – valores da temperatura do ar estimados a partir da temperatura do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 9** - Comparativo entre os valores reais (azul) e os estimados a partir das temperaturas



Fonte: Do autor (2023)

Analisando a relação entre a temperatura do ar (Bulbo Seco) como variável de entrada e a previsão da temperatura do ponto de orvalho com um intervalo de sete dias como variável de saída (Figuras 10 e 11) (Tabela 6). O modelo SVM apresentou um NSE negativo (-11,042), indicando um desempenho substancialmente inferior em relação aos demais modelos testados, situação comparável apenas com o pior caso de Dayal *et al.* (2021). Seus coeficientes de correlação ( $d=0,331$ ,  $r=0,678$ ,  $c=0,224$ ,  $KGE=-0,251$ ) sugerem uma relação limitada entre as variáveis, enquanto o RMSE de

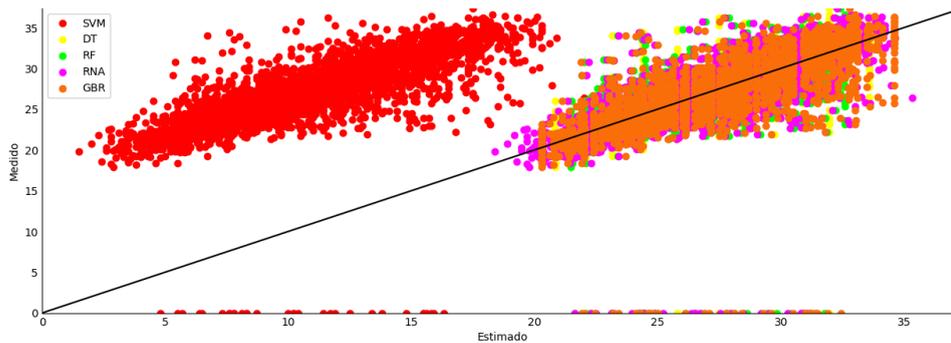
16,367 aponta para uma precisão insatisfatória na previsão. Em contraste, os modelos DT, RF, GBR e ANN demonstraram NSE variando de 0,447 a 0,460, com coeficientes de correlação superiores a 0,67 e RMSE na faixa de 3,467 a 3,509. Na combinação, o modelo de destaque foi o GBR, assim como em Xu *et al.* (2023), esse modelo apresentou uma maior velocidade para o processamento e para a otimização, mesmo utilizando parâmetros hidrológicos diferentes em cada estudo.

**Tabela 6** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE     | d     | r     | c     | KGE    | R <sup>2</sup> | RMSE   | SM     |
|--------|---------|-------|-------|-------|--------|----------------|--------|--------|
| SVM    | -11.042 | 0.331 | 0.678 | 0.224 | -0.251 | 0.460          | 16.367 | 31.967 |
| DT     | 0.454   | 0.787 | 0.675 | 0.531 | 0.551  | 0.455          | 3.484  | 6.032  |
| RF     | 0.458   | 0.788 | 0.677 | 0.534 | 0.551  | 0.459          | 3.473  | 6.006  |
| GBR    | 0.447   | 0.795 | 0.675 | 0.537 | 0.578  | 0.456          | 3.509  | 6.021  |
| ANN    | 0.460   | 0.788 | 0.679 | 0.535 | 0.553  | 0.461          | 3.467  | 5.992  |

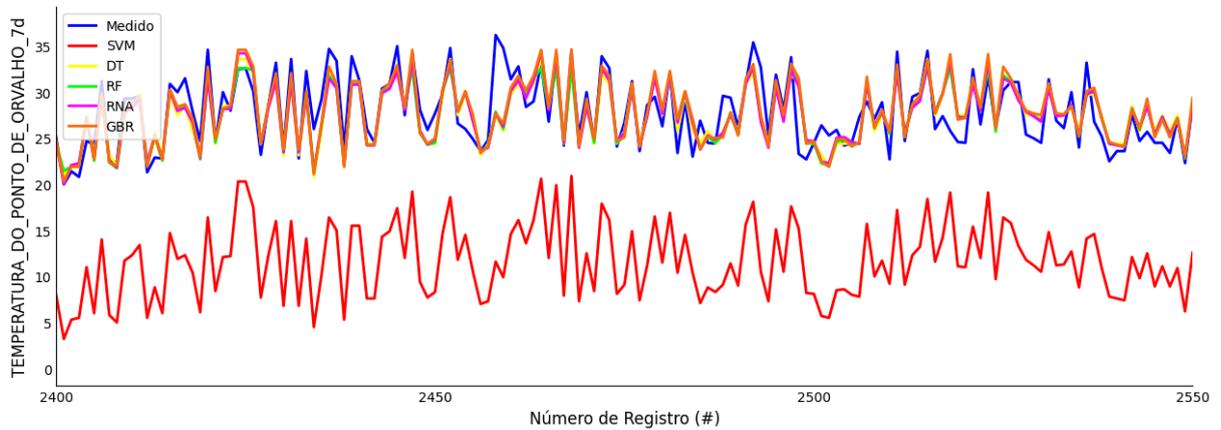
Fonte: Autor (2024)

**Figura 10** - Dispersão dos resultados previstos – valores da temperatura no ponto de orvalho estimados a partir da temperatura do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 11** - Comparativo entre os valores reais (azul) e os estimados a partir das temperaturas



Fonte: Do autor (2023)

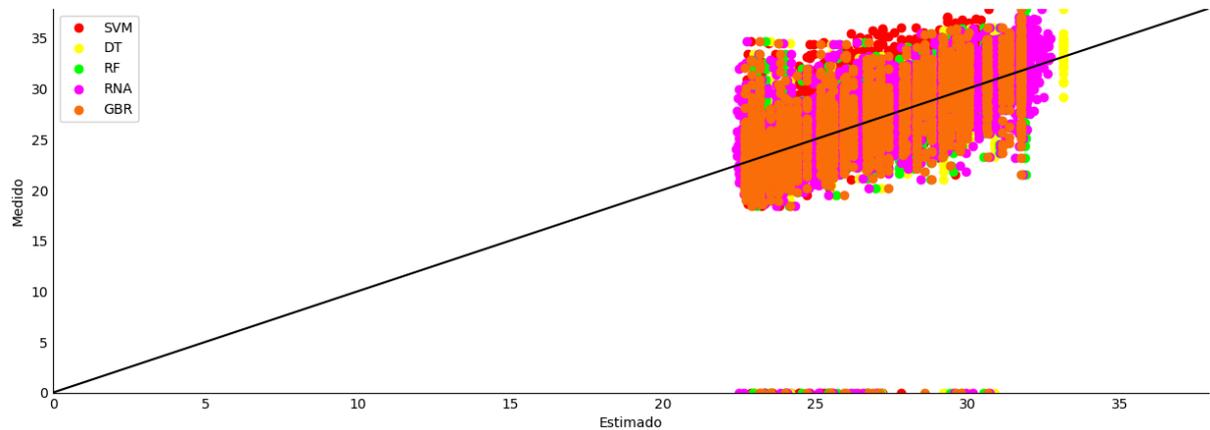
Analisando a relação entre a umidade relativa do ar como variável de entrada e a temperatura do ar (bulbo seco) com um intervalo de 3 dias como variável de saída (Figuras 12 e 13) (Tabela 7). O modelo SVM apresentou um NSE de 0,309, com coeficientes de correlação ( $d=0,672$ ,  $r=0,682$ ,  $c=0,459$ ,  $KGE=0,390$ ) sugerindo uma relação moderada entre as variáveis, distante do encontrado no estudo de Lal e Datta (2019). No entanto, seu RMSE de 3,703 indica uma precisão limitada na previsão. Por outro lado, os modelos DT, RF, GBR e ANN demonstraram NSE variando de 0,463 a 0,470, com coeficientes de correlação acima de 0,68 e RMSE na faixa de 3,244 à 3,265.

**Tabela 7** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE   | d     | r     | c     | KGE   | R <sup>2</sup> | RMSE  | SM    |
|--------|-------|-------|-------|-------|-------|----------------|-------|-------|
| SVM    | 0.309 | 0.672 | 0.682 | 0.459 | 0.390 | 0.465          | 3.703 | 6.725 |
| DT     | 0.467 | 0.786 | 0.684 | 0.538 | 0.540 | 0.468          | 3.252 | 5.769 |
| RF     | 0.470 | 0.786 | 0.686 | 0.540 | 0.539 | 0.471          | 3.244 | 5.752 |
| GBR    | 0.463 | 0.783 | 0.681 | 0.533 | 0.537 | 0.464          | 3.265 | 5.803 |
| ANN    | 0.468 | 0.785 | 0.684 | 0.538 | 0.538 | 0.468          | 3.251 | 5.769 |

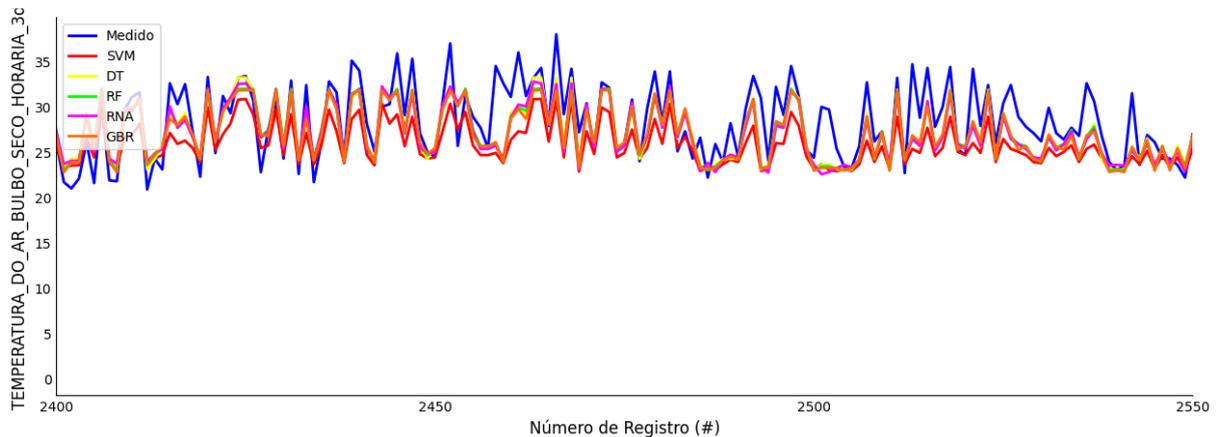
Fonte: Autor (2024)

**Figura 12** - Dispersão dos resultados previstos – valores da umidade do ar estimados a partir da temperatura do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 13** - Comparativo entre os valores reais (azul) e os estimados a partir da temperatura e da umidade do ar



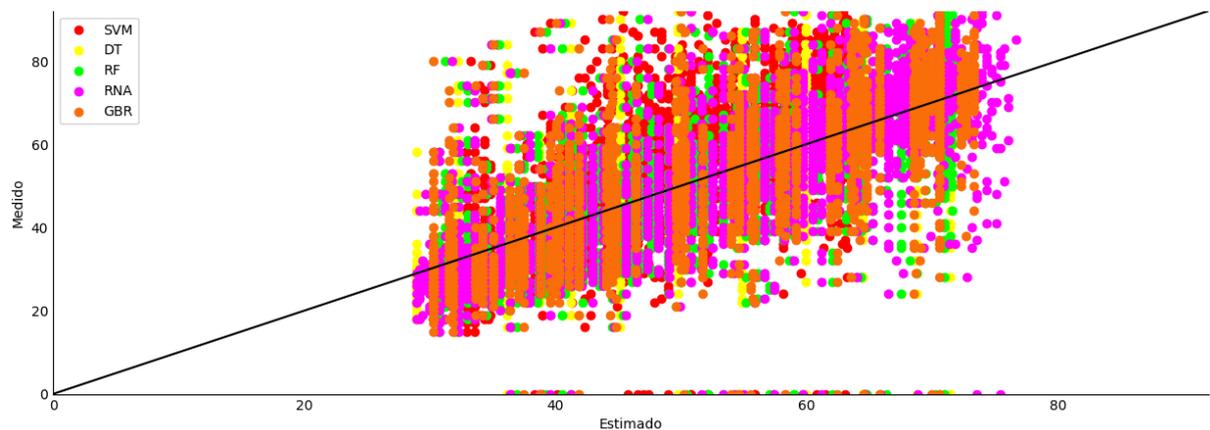
Fonte: Do autor (2023)

Ao examinar a relação entre a umidade relativa do ar como variável de entrada e a previsão da umidade relativa do ar com um intervalo de três dias como variável de saída (Figuras 14 e 15) (Tabela 8). O modelo SVM demonstrou um NSE de 0,409, com coeficientes de correlação ( $d=0,744$ ,  $r=0,744$ ,  $c=0,554$ ,  $KGE=0,489$ ) indicando uma relação moderada entre as variáveis. No entanto, seu RMSE de 13,622 sugere uma precisão limitada na previsão. Em contraste, os modelos DT, RF, GBR e ANN apresentaram NSE variando de 0,556 a 0,561, com coeficientes de correlação acima de 0,74 e RMSE na faixa de 11,743 a 11,808. Interessante notar que para algoritmos e combinações com bons resultados, o NSE para RF desempenha um pouco melhor do que em outros modelos, assim como em Liang *et al.* (2018).

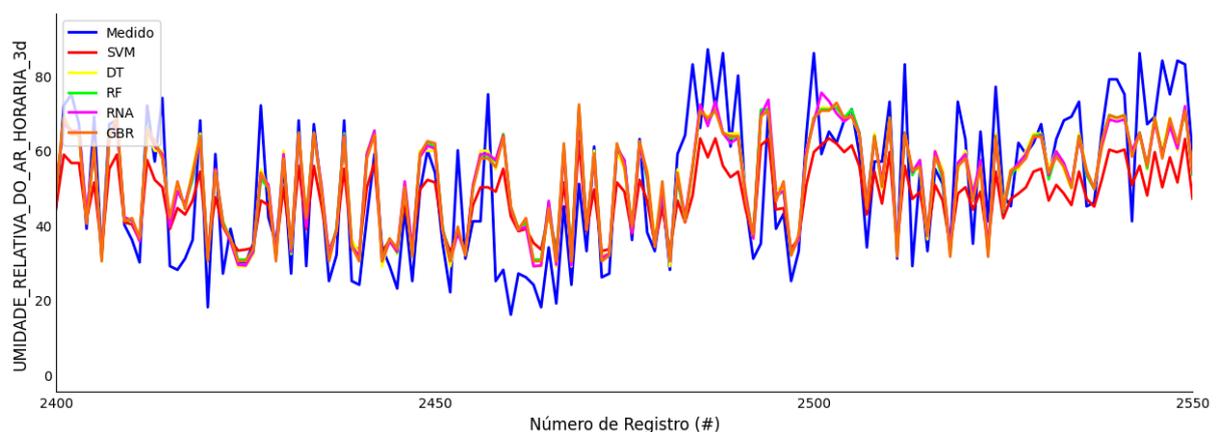
**Tabela 8** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE   | d     | r     | c     | KGE   | R <sup>2</sup> | RMSE   | SM     |
|--------|-------|-------|-------|-------|-------|----------------|--------|--------|
| SVM    | 0.409 | 0.744 | 0.744 | 0.554 | 0.489 | 0.554          | 13.622 | 16.127 |
| DT     | 0.556 | 0.841 | 0.746 | 0.627 | 0.628 | 0.557          | 11.807 | 13.853 |
| RF     | 0.561 | 0.843 | 0.749 | 0.631 | 0.631 | 0.562          | 11.743 | 13.766 |
| GBR    | 0.556 | 0.842 | 0.746 | 0.628 | 0.636 | 0.556          | 11.808 | 13.842 |
| ANN    | 0.559 | 0.841 | 0.748 | 0.629 | 0.628 | 0.559          | 11.776 | 13.813 |

Fonte: Autor (2024)

**Figura 14** - Dispersão dos resultados previstos – valores da umidade do ar estimados a partir da umidade do ar versus os valores medidos (reais).

Fonte: Do autor (2023)

**Figura 15** - Comparativo entre os valores reais (azul) e os estimados a partir das umidades

Fonte: Do autor (2023)

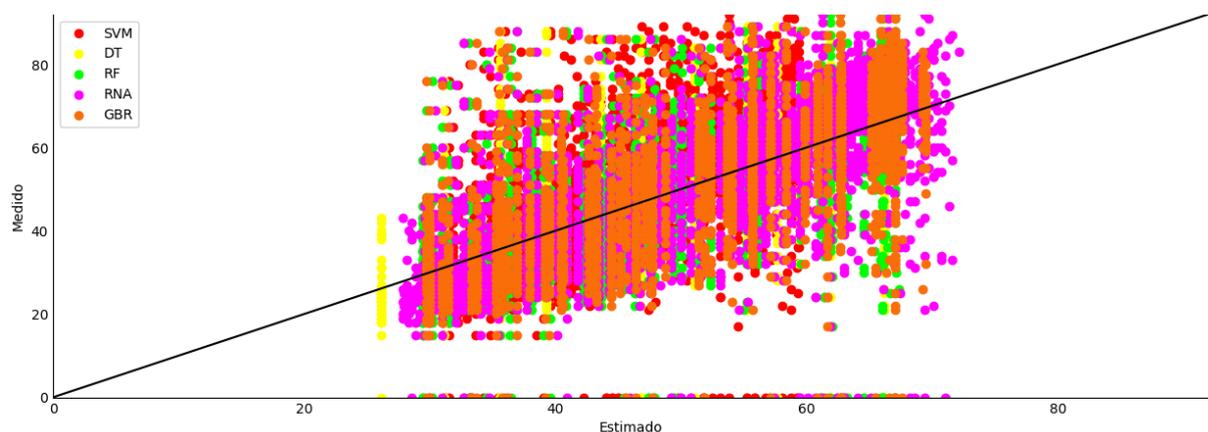
Ao investigar a relação da umidade relativa do ar como variável de entrada com a previsão da umidade relativa do ar com um intervalo de sete dias como variável de saída (Figuras 16 e 17) (Tabela 9). O modelo SVM registrou um NSE de 0,346, com coeficientes de correlação ( $d=0,708$ ,  $r=0,687$ ,  $c=0,487$ ,  $KGE=0,448$ ) indicando uma relação moderada entre as variáveis. No entanto, seu RMSE de 14,328 sugere uma precisão limitada na previsão. Em contraste, os modelos DT, RF, GBR e ANN exibiram NSE variando de 0,469 a 0,475, com coeficientes de correlação acima de 0,68 e RMSE na faixa de 12,844 a 12,909.

**Tabela 9** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE   | d     | r     | c     | KGE   | R <sup>2</sup> | RMSE   | SM     |
|--------|-------|-------|-------|-------|-------|----------------|--------|--------|
| SVM    | 0.346 | 0.708 | 0.687 | 0.487 | 0.448 | 0.472          | 14.328 | 17.180 |
| DT     | 0.472 | 0.799 | 0.687 | 0.549 | 0.560 | 0.473          | 12.881 | 15.341 |
| RF     | 0.475 | 0.801 | 0.690 | 0.553 | 0.564 | 0.476          | 12.844 | 15.286 |
| GBR    | 0.470 | 0.801 | 0.687 | 0.550 | 0.567 | 0.471          | 12.895 | 15.349 |
| ANN    | 0.469 | 0.797 | 0.686 | 0.547 | 0.557 | 0.470          | 12.909 | 15.383 |

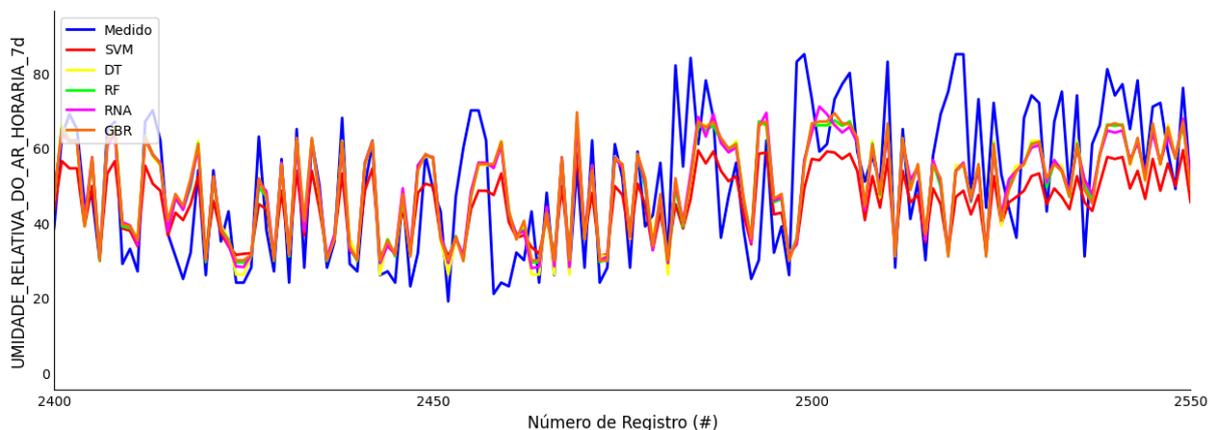
Fonte: Autor (2024)

**Figura 16** - Dispersão dos resultados previstos – valores da umidade do ar estimados a partir da umidade do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 17** - Comparativo entre os valores reais (azul) e os estimados a partir das umidades do ar



Fonte: Do autor (2023)

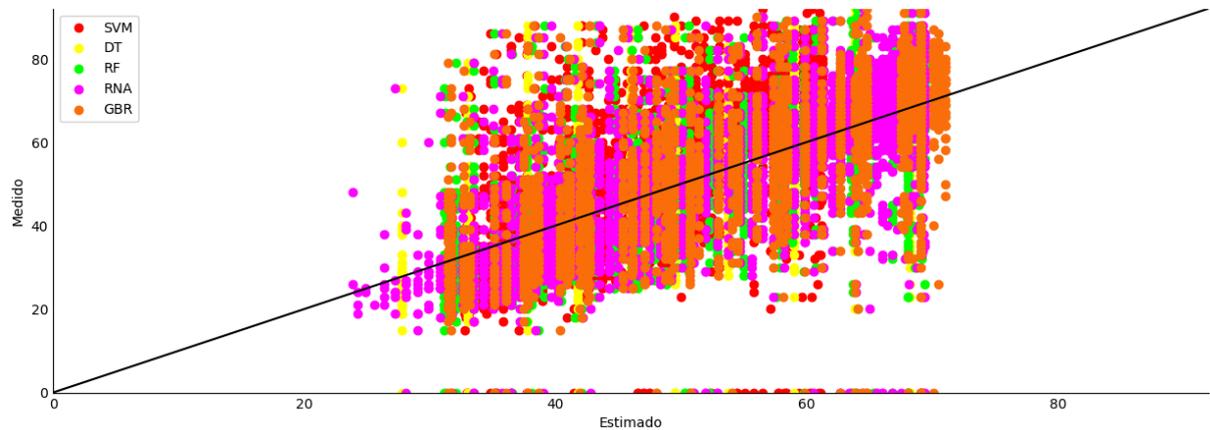
Ao examinar a relação entre a umidade relativa do ar como variável de entrada e a velocidade do vento com um intervalo de sete dias como variável de saída (Figuras 18 e 19) (Tabela 10). O modelo SVM registrou um NSE de 0,346, indicando uma relação moderada entre as variáveis, com coeficientes de correlação ( $d=0,708$ ,  $r=0,683$ ,  $c=0,483$ ,  $KGE=0,444$ ). No entanto, seu RMSE de 14,392 sugere uma precisão limitada na previsão. Em contrapartida, os modelos DT, RF, GBR e ANN exibiram NSE variando de 0,463 a 0,467, com coeficientes de correlação acima de 0,68 e RMSE na faixa de 12,993 a 13,044.

**Tabela 10** - Resultado das métricas para a combinação mencionada.

| Modelo | NSE   | d     | r     | c     | KGE   | R <sup>2</sup> | RMSE   | SM     |
|--------|-------|-------|-------|-------|-------|----------------|--------|--------|
| SVM    | 0.346 | 0.708 | 0.683 | 0.483 | 0.444 | 0.466          | 14.392 | 17.263 |
| DT     | 0.463 | 0.795 | 0.681 | 0.541 | 0.553 | 0.463          | 13.044 | 15.548 |
| RF     | 0.467 | 0.798 | 0.684 | 0.546 | 0.558 | 0.468          | 12.993 | 15.473 |
| GBR    | 0.463 | 0.797 | 0.681 | 0.543 | 0.559 | 0.463          | 13.043 | 15.537 |
| ANN    | 0.466 | 0.797 | 0.683 | 0.544 | 0.554 | 0.467          | 13.003 | 15.493 |

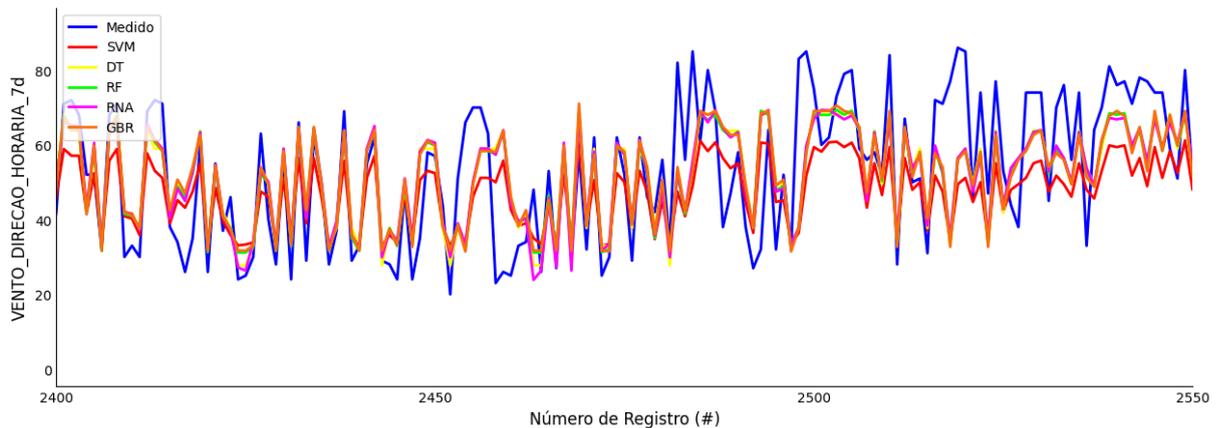
Fonte: Autor (2024)

**Figura 18** - Dispersão dos resultados previstos – valores da velocidade do vento estimados a partir da umidade do ar versus os valores medidos (reais).



Fonte: Do autor (2023)

**Figura 19** - Comparativo entre os valores reais (azul) e os estimados a partir da velocidade e da umidade do ar



Fonte: Do autor (2023)

Após realizadas as análises para diversos modelos e variáveis de entrada e saída, é importante destacar que para as demais combinações testadas, não foram identificadas informações relevantes nos resultados. Portanto, para esses casos não foi realizada a geração de gráficos ou uma análise mais detalhada do desempenho, provenientes dos diferentes processamentos e modelos.

Considerando todos os resultados encontrados expostos no Apêndice D, foram identificadas combinações envolvendo temperatura do ar e do ponto de orvalho, velocidade do vento e umidade relativa do ar que conseguiram gerar um desempenho suficiente para serem consideradas em aplicações práticas utilizando os algoritmos desenvolvidos. Infelizmente, resultados com bons desempenho para a

precipitação não foram obtidos, diferentemente de Jimeno Sáez *et al.* (2018) onde a maioria de seus modelos com melhores resultados foram de precipitação e vazão. Especificamente, a combinação de umidade relativa do ar como informação de entrada e temperatura do ar como informação de saída foi responsável por dez dos dez melhores resultados dos algoritmos.

Ao avaliar os valores das métricas, constatou-se que o algoritmo de *random forest* (RF) obteve o melhor desempenho. No entanto, ao considerar o número de ocorrências em que cada algoritmo teve um bom resultado, o gradiente *boosting regressor* e as redes neurais artificiais emergiram como vencedores do título de melhor algoritmo para o estudo.

---

## 4 CONCLUSÕES

Com base nas análises realizadas utilizando diversos modelos e combinações de variáveis de entrada e saída, torna-se evidente que determinadas configurações se destacaram, proporcionando insights valiosos sobre a relação entre as variáveis climáticas estudadas.

Ao considerar a temperatura do ar como variável de entrada e a umidade relativa do ar como saída, especialmente no intervalo temporal de 0 dias, os resultados obtidos por diferentes algoritmos demonstraram consistência e desempenho satisfatório, exceto para o algoritmo SVM, que apresentou um desempenho inferior. No entanto, para outras combinações de variáveis, como a relação entre a temperatura do ar e a umidade relativa do ar com intervalos maiores, os modelos DT, RF, GBR e ANN se destacaram, exibindo resultados mais precisos e consistentes, enquanto o SVM teve um desempenho inferior.

É relevante observar que, embora algumas relações tenham sido identificadas como moderadas ou apresentaram desempenho inferior, isso proporcionou uma compreensão mais profunda das interações entre as variáveis climáticas estudadas. Modelos como DT, RF, GBR e ANN mostraram uma capacidade mais robusta de previsão e precisão em diversas combinações de variáveis, evidenciando sua aplicabilidade e confiabilidade para esses contextos específicos.

---

## REFERÊNCIAS

ALANEME, Uche George; MBADIKE, Elvis Mbanefo. *Modelling of the mechanical properties of concrete with cement ratio partially replaced by aluminium waste and sawdust ash using artificial neural network*. **SN Applied Sciences**, v. 1, n. 11, p. 1514, 2019. Springer.

ANTONIADIS, Anestis; LAMBERT-LACROIX, Sophie; POGGI, Jean-Michel. *Random forests for global sensitivity analysis: A selective review*. **Reliability Engineering & System Safety**, v. 206, p. 107312, 2021. Elsevier.

ARAUNA, Kevin Victor. **Comparison between Stochastic Gradient Descent and Support Vector Machine on eCommerce Data using RandomizedSearchCV**. PhD thesis. Universitas Katholik Soegijapranata Semarang, 2023.

AZMAIN, Mahfuz Ahmed; TAHMID, Ahnaf; CHOWDHURY, Mooaj Sadaqat; HOSSAM-E-HAIDER, Md. *Prediction of Rainfall in Bangladesh: A Case Study of the Machine Learning*. **2022 IEEE 7th International conference for Convergence in Technology (I2CT)**, pp. 1-5. IEEE, 2022.

BARBOSA, Wenderson Gomes; DE MOURA CARVALHO, Jaine; DA SILVA, Denise Honorato Lopes; DA SILVA JÚNIOR, Arundo Nunes; DA SILVA ARAÚJO, Lidiane; DA SILVA, Antonio Samuel Alves; FERREIRA, Tiago Alessandro Espínola; CRISTINO, Claudio Tadeu; STOSIC, Tatijana. Sazonalidade do regime de chuva nas mesorregiões do estado de Pernambuco, Brasil. **Research, Society and Development**, v. 12, n. 12, 2023.

BRAZ, Jean Alexsander Gomes. *Predição de velocidade de ventos para geração eólica, utilizando aprendizado de máquina*. Trabalho de Conclusão de Curso. Florianópolis, SC., 2023.

CHEN, Rung-Ching; DEWI, Christine; HUANG, Su-Wen; CARAKA, Rezzy Eko. *Selecting critical features for data classification based on machine learning methods*. **Journal of Big Data**, v. 7, n. 1, p. 52, 2020. Springer.

COHEN, Israel; HUANG, Yiteng; CHEN, Jingdong; BENESTY, Jacob. Pearson correlation coefficient. **Noise reduction in speech processing**, p. 1-4, 2009. Springer.

DALLA TORRE, Daniele; LOMBARDI, Andrea; MENAPACE, Andrea; ZANFEI, Ariele; RIGHETTI, Maurizio. *Exploring the feasibility of Support Vector Machine for short-term hydrological forecasting in South Tyrol: challenges and prospects*. **Discover Applied Sciences**, v. 6, n. 4, p. 154, 2024. Springer.

DAYAL, Deen; GUPTA, Praveen K; PANDEY, Ashish. *Streamflow estimation using satellite-retrieved water fluxes and machine learning technique over monsoon-dominated catchments of India*. **Hydrological Sciences Journal**, v. 66, n. 4, p. 656-671, 2021. Taylor & Francis.

DA SILVA, Rodrigo Dlugosz; DE AGUIAR, Marcelo Augusto; CANTERI, Marcelo Giovanetti; ROSISCA, Juliandra Rodrigues; JUNIO, Nilson Aparecido Vieira. *Reference evapotranspiration for Londrina, Paraná, Brazil: Performance of different*

*estimation methods. Semina: Ciências Agrárias*, v. 38, n. 4, p. 2363-2374, 2017. Universidade Estadual de Londrina.

DA SILVA, Bruno Fonseca; DOS SANTOS RODRIGUES, Rodrigo Zimmerle; HEISKANEN, Janne; ABERA, Temesgen Alemayehu; GASPARETTO, Suelen Cristina; BIASE, Adriele Giaretta; BALLESTER, Maria Victoria Ramos; DE MOURA, Yhasmin Mendes; DE STEFANO PIEDADE, Sônia Maria; DE OLIVEIRA SILVA, Andrezza Karla. *Evaluating the temporal patterns of land use and precipitation under desertification in the semi-arid region of Brazil. Ecological Informatics*, v. 77, p. 102192, 2023. Elsevier.

DUMITRESCU, Elena; HUÉ, Sullivan; HURLIN, Christophe; TOKPAVI, Sessi. *Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research*, v. 297, n. 3, p. 1178-1192, 2022. Elsevier.

FURQUIM, Gustavo; PESSIN, Gustavo; FAIÇAL, Bruno S; MENDIONDO, Eduardo M; UEYAMA, Jó. *Improving the accuracy of a flood forecasting model by means of machine learning and chaos theory: A case study involving a real wireless sensor network deployment in Brazil. Neural Computing and Applications*, v. 27, p. 1129-1141, 2016. Springer.

HUANG, Huaping; LIANG, Zhongmin; LI, Binqun; WANG, Dong; HU, Yiming; LI, Yujie. *Combination of multiple data-driven models for long-term monthly runoff predictions based on Bayesian model averaging. Water Resources Management*, v. 33, p. 3321-3338, 2019. Springer.

JIMENO SÁEZ, Patricia et al. *Simulación de procesos hidrológicos utilizando técnicas de machine learning y modelos hidrológicos*. 2018.

KAMIR, Elisa; WALDNER, François; HOCHMAN, Zvi. *Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS Journal of Photogrammetry and Remote Sensing*, v. 160, p. 124-135, 2020. Elsevier.

LAL, Alvin; DATTA, Bithin. *Multi-objective groundwater management strategy under uncertainties for sustainable control of saltwater intrusion: Solution for an island country in the South Pacific. Journal of Environmental Management*, v. 234, p. 115-130, 2019. Elsevier.

LATIF, Sarmad Dashti; AHMED, Ali Najah. *A review of deep learning and machine learning techniques for hydrological inflow forecasting. Environment, Development and Sustainability*, v. 25, n. 11, p. 12189-12216, 2023. Springer.

LIAKOS, Konstantinos G; BUSATO, Patrizia; MOSHOU, Dimitrios; PEARSON, Simon; BOCHTIS, Dionysis. *Machine learning in agriculture: A review. Sensors*, v. 18, n. 8, p. 2674, 2018. Mdpi.

LIANG, Zhongmin; TANG, Tiantian; LI, Binqun; LIU, Tian; WANG, Jun; HU, Yiming. *Long-term streamflow forecasting using SWAT through the integration of the random*

*forests precipitation generator: case study of Danjiangkou Reservoir. Hydrology Research*, v. 49, n. 5, p. 1513-1527, 2018. IWA Publishing.

MORAES, Rafael Guimarães Silva; LIMA, Elton Ferreira; DE SOUSA OLIVEIRA, Pedro Luid; DAMASCENA, Jossimara Ferreira; DA SILVA, Cristiane Matos. Métodos de estimativa da evapotranspiração de referência no período seco e chuvoso em Imperatriz, MA. **Revista Brasileira de Climatologia**, v. 33, p. 169-188, 2023.

NACHAPPA, Thimmaiah Gudiyangada; PIRALILLOU, Sepideh Tavakkoli; GHOLAMNIA, Khalil; GHORBANZADEH, Omid; RAHMATI, Omid; BLASCHKE, Thomas. *Flood susceptibility mapping with machine learning, multi-criteria decision analysis and ensemble using Dempster Shafer Theory. Journal of Hydrology*, v. 590, p. 125275, 2020. Elsevier.

NGUYEN, Thieu; NGUYEN, Binh Minh; NGUYEN, Giang. *Building resource auto-scaler with functional-link neural network and adaptive bacterial foraging optimization. International Conference on Theory and Applications of Models of Computation*, p. 501-517, 2019. Springer.

PRIETO, Cristina; LE VINE, Nataliya; KAVETSKI, Dmitri; GARCÍA, Eduardo; MEDINA, Raúl. *Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests. Water Resources Research*, v. 55, n. 5, p. 4364-4392, 2019. Wiley Online Library.

RAJAMANI, Santhosh Kumar; IYER, Radha Srinivasan. *Machine Learning-Based Mobile Applications Using Python and Scikit-Learn. Designing and developing innovative mobile applications*, p. 282-306, 2023. IGI Global.

SILVA, Dao; SILVA, Jlb; OLIVEIRA, Jda; BATISTA, Phd; LOPES, Pmo; MOURA, Gba. Evolução espaço-temporal da vegetação em região semiárida. **Geotecnologias Aplicada a Estudos Ambientais**. EPGRAF, Campina Grande, p. 87-101, 2018.

VAN THIEU, Nguyen; BARMA, Surajit Deb; VAN LAM, To; KISI, Ozgur; MAHESHA, Amai. *Groundwater level modeling using augmented artificial ecosystem optimization. Journal of Hydrology*, v. 617, p. 129034, 2023. Elsevier.

VISHNU, MK; RUPAK, VR Vishal; VEDHAPRIYAA, S; SANGEETHA, M; MANJULADEVI, R; SAGANA, C. *Recurrent gastric cancer prediction using randomized search cv optimizer. 2023 International Conference on Computer Communication and Informatics (ICCCI)*, p. 1-5, 2023. IEEE.

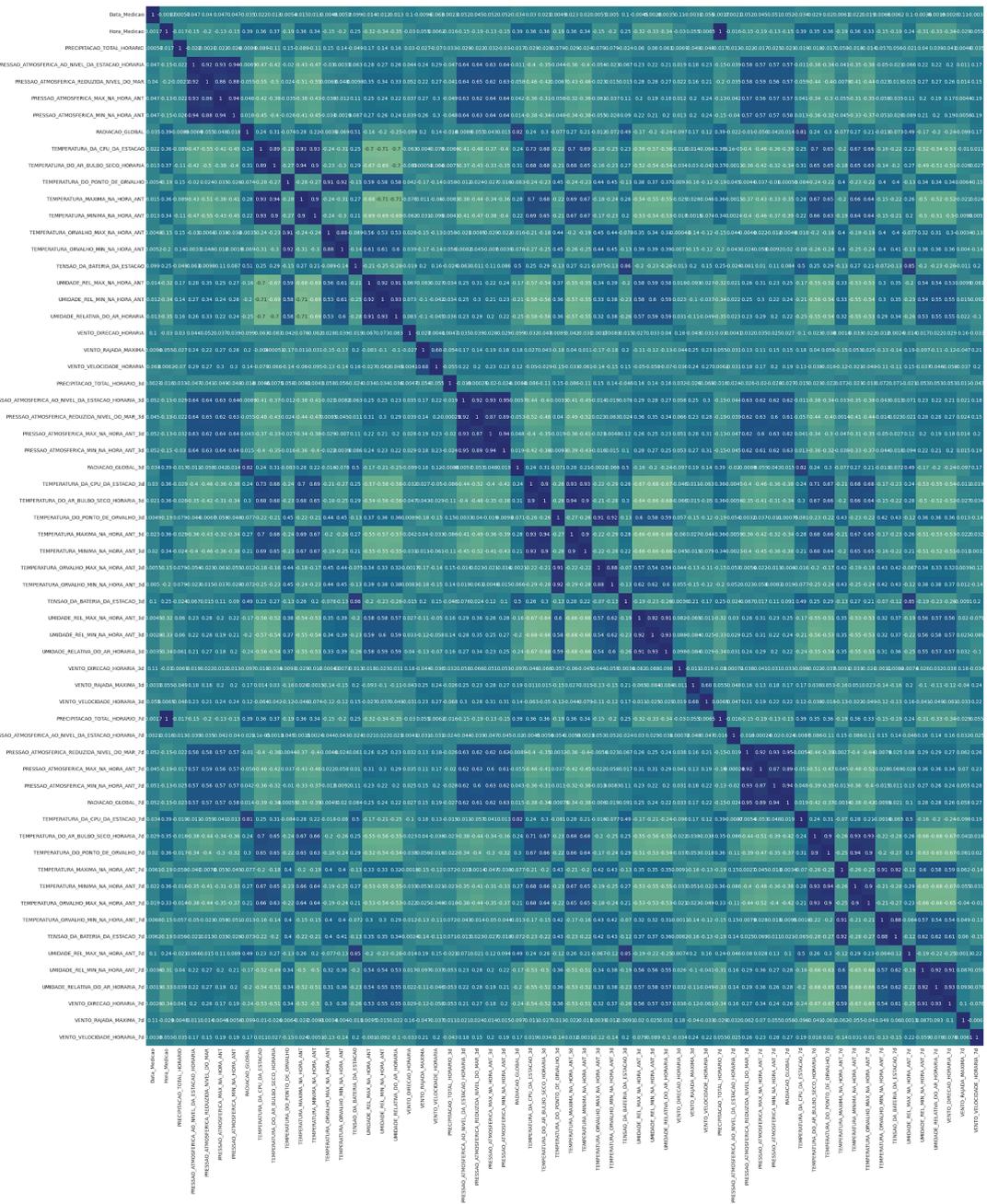
XU, Kui; HAN, Zhentao; XU, Hongshi; BIN, Lingling. *Rapid prediction model for urban floods based on a light gradient boosting machine approach and hydrological-hydraulic model. International Journal of Disaster Risk Science*, v. 14, n. 1, p. 79-97, 2023. Springer.

YU, Han; HUTSON, Alan D. *A robust Spearman correlation coefficient permutation test. Communications in Statistics-Theory and Methods*, v. 53, n. 6, p. 2141-2153, 2024. Taylor & Francis.

ZADOR, Anthony M. *A critique of pure learning and what artificial neural networks can learn from animal brains*. **Nature Communications**, v. 10, n. 1, p. 3770, 2019. Nature Publishing Group UK London.



### Apêndice B - Gráfico da matriz de correlação pelo método de Kendall, gerado pelo código.







|   |   |                               |  |   |   |   |  |   |  |   |
|---|---|-------------------------------|--|---|---|---|--|---|--|---|
| O_HORARI<br>A   |   | ANN                           | 0.376                                      | 0.738                                     | 0.614                                     | 0.453                                     | 0.461                                      | 0.377                                     | 3.072  | 6.053                                       |
| PRESSAO<br>_ATMOSF<br>ERICA_AO<br>_NIVEL_D<br>A_ESTACA<br>O_HORARI<br>A | UMIDADE_<br>RELATIVA<br>_DO_AR_<br>HORARIA                  | SVM<br>DT<br>RF<br>GBR<br>ANN | -0.004<br>0.136<br>0.139<br>0.124<br>0.129 | 0.092<br>0.503<br>0.503<br>0.499<br>0.489 | nan<br>0.373<br>0.377<br>0.359<br>0.362   | nan<br>0.188<br>0.189<br>0.179<br>0.177   | nan<br>0.145<br>0.146<br>0.131<br>0.115    | nan<br>0.139<br>0.142<br>0.129<br>0.131   | 17.008<br>15.783<br>15.757<br>15.895<br>15.842 | nan<br>20.298<br>20.262<br>20.474<br>20.438 |
| PRESSAO<br>_ATMOSF<br>ERICA_AO<br>_NIVEL_D<br>A_ESTACA<br>O_HORARI<br>A | VENTO_VE<br>LOCIDADE<br>_HORARIA                            | SVM<br>DT<br>RF<br>GBR<br>ANN | 0.000<br>0.166<br>0.170<br>0.162<br>0.174  | 0.005<br>0.535<br>0.537<br>0.543<br>0.538 | 0.000<br>0.409<br>0.413<br>0.406<br>0.418 | 0.000<br>0.219<br>0.222<br>0.221<br>0.225 | -0.414<br>0.187<br>0.188<br>0.187<br>0.192 | 0.000<br>0.168<br>0.171<br>0.165<br>0.174 | 1.353<br>1.235<br>1.232<br>1.239<br>1.230      | 7.762<br>5.551<br>5.532<br>5.556<br>5.509   |
| PRESSAO<br>_ATMOSF<br>ERICA_AO<br>_NIVEL_D<br>A_ESTACA<br>O_HORARI      | TEMPERA<br>TURA_DO<br>_AR_BUL<br>BO_SECO<br>_HORARIA<br>_3d | SVM<br>DT<br>RF<br>GBR        | -0.001<br>0.215<br>0.218<br>0.204          | 0.035<br>0.578<br>0.578<br>0.577          | nan<br>0.464<br>0.467<br>0.452            | nan<br>0.268<br>0.270<br>0.261            | nan<br>0.240<br>0.242<br>0.241             | nan<br>0.215<br>0.218<br>0.204            | 4.457<br>3.949<br>3.940<br>3.976               | nan<br>7.970<br>7.946<br>8.036              |



|          |          |     |        |       |       |       |        |       |        |        |
|----------|----------|-----|--------|-------|-------|-------|--------|-------|--------|--------|
| PRESSAO  | UMIDADE_ | SVM | -0.006 | 0.112 | nan   | nan   | nan    | nan   | 17.770 | nan    |
| _ATMOSF  | RELATIVA |     |        |       |       |       |        |       |        |        |
| ERICA_AO | _DO_AR_  | DT  | 0.088  | 0.424 | 0.305 | 0.129 | 0.044  | 0.093 | 16.919 | 21.836 |
| _NIVEL_D | HORARIA_ | RF  | 0.092  | 0.426 | 0.312 | 0.133 | 0.047  | 0.097 | 16.882 | 21.776 |
| A_ESTACA | 7d       |     |        |       |       |       |        |       |        |        |
| O_HORARI |          | GBR | 0.074  | 0.436 | 0.289 | 0.126 | 0.051  | 0.084 | 17.050 | 21.991 |
| A        |          |     |        |       |       |       |        |       |        |        |
|          |          | ANN | 0.085  | 0.409 | 0.298 | 0.122 | 0.018  | 0.089 | 16.951 | 21.931 |
| PRESSAO  | VENTO_VE | SVM | -0.001 | 0.052 | 0.000 | 0.000 | -0.414 | 0.000 | 2.456  | 8.820  |
| _ATMOSF  | LOCIDADE |     |        |       |       |       |        |       |        |        |
| ERICA_AO | _HORARIA | DT  | 0.066  | 0.345 | 0.258 | 0.089 | -0.044 | 0.066 | 2.372  | 7.592  |
| _NIVEL_D | _7d      | RF  | 0.069  | 0.350 | 0.263 | 0.092 | -0.041 | 0.069 | 2.368  | 7.566  |
| A_ESTACA |          |     |        |       |       |       |        |       |        |        |
| O_HORARI |          | GBR | 0.054  | 0.352 | 0.238 | 0.084 | -0.043 | 0.057 | 2.388  | 7.647  |
| A        |          |     |        |       |       |       |        |       |        |        |
|          |          | ANN | 0.067  | 0.339 | 0.259 | 0.088 | -0.044 | 0.067 | 2.371  | 7.596  |
| TEMPERA  | UMIDADE_ | SVM | 0.408  | 0.833 | 0.877 | 0.730 | 0.774  | 0.770 | 13.059 | 14.667 |
| TURA_DO  | RELATIVA |     |        |       |       |       |        |       |        |        |
| _AR_BUL  | _DO_AR_  | DT  | 0.774  | 0.934 | 0.880 | 0.822 | 0.838  | 0.775 | 8.063  | 9.040  |
| BO_SECO  | HORARIA  | RF  | 0.780  | 0.935 | 0.884 | 0.826 | 0.842  | 0.781 | 7.966  | 8.918  |
| _HORARIA |          |     |        |       |       |       |        |       |        |        |
|          |          | GBR | 0.775  | 0.934 | 0.881 | 0.823 | 0.840  | 0.777 | 8.048  | 9.019  |
|          |          |     |        |       |       |       |        |       |        |        |
|          |          | ANN | 0.770  | 0.932 | 0.878 | 0.818 | 0.834  | 0.770 | 8.146  | 9.145  |
| TEMPERA  | TEMPERA  | SVM | 0.072  | 0.699 | 0.765 | 0.535 | 0.595  | 0.585 | 4.293  | 7.043  |





|                 |                 |            |              |              |              |              |               |              |               |               |
|-----------------|-----------------|------------|--------------|--------------|--------------|--------------|---------------|--------------|---------------|---------------|
| <b>_HORARIA</b> |                 | <b>GBR</b> | <b>0.403</b> | <b>0.764</b> | <b>0.637</b> | <b>0.487</b> | <b>0.513</b>  | <b>0.405</b> | <b>13.748</b> | <b>16.539</b> |
|                 |                 | <b>ANN</b> | <b>0.412</b> | <b>0.765</b> | <b>0.644</b> | <b>0.492</b> | <b>0.414</b>  | <b>0.414</b> | <b>13.643</b> | <b>16.402</b> |
| <b>UMIDADE_</b> | <b>PRESSAO</b>  | <b>SVM</b> | <b>0.000</b> | <b>0.020</b> | <b>0.000</b> | <b>0.000</b> | <b>-0.414</b> | <b>0.000</b> | <b>2.919</b>  | <b>9.313</b>  |
| <b>RELATIVA</b> | <b>_ATMOSF</b>  |            |              |              |              |              |               |              |               |               |
| <b>_DO_AR_</b>  | <b>ERICA_AO</b> | <b>DT</b>  | <b>0.184</b> | <b>0.542</b> | <b>0.429</b> | <b>0.233</b> | <b>0.189</b>  | <b>0.184</b> | <b>2.637</b>  | <b>6.876</b>  |
| <b>HORARIA</b>  | <b>_NIVEL_D</b> | <b>RF</b>  | <b>0.185</b> | <b>0.537</b> | <b>0.431</b> | <b>0.231</b> | <b>0.186</b>  | <b>0.186</b> | <b>2.635</b>  | <b>6.880</b>  |
|                 | <b>A_ESTACA</b> |            |              |              |              |              |               |              |               |               |
|                 | <b>O_HORARI</b> | <b>GBR</b> | <b>0.173</b> | <b>0.541</b> | <b>0.417</b> | <b>0.226</b> | <b>0.189</b>  | <b>0.174</b> | <b>2.655</b>  | <b>6.935</b>  |
|                 | <b>A</b>        |            |              |              |              |              |               |              |               |               |
|                 |                 | <b>ANN</b> | <b>0.184</b> | <b>0.539</b> | <b>0.430</b> | <b>0.232</b> | <b>0.182</b>  | <b>0.185</b> | <b>2.636</b>  | <b>6.884</b>  |
| <b>UMIDADE_</b> | <b>TEMPERA</b>  | <b>SVM</b> | <b>0.765</b> | <b>0.926</b> | <b>0.877</b> | <b>0.812</b> | <b>0.790</b>  | <b>0.770</b> | <b>1.884</b>  | <b>2.943</b>  |
| <b>RELATIVA</b> | <b>TURA_DO</b>  |            |              |              |              |              |               |              |               |               |
| <b>_DO_AR_</b>  | <b>_AR_BUL</b>  | <b>DT</b>  | <b>0.784</b> | <b>0.936</b> | <b>0.886</b> | <b>0.829</b> | <b>0.834</b>  | <b>0.784</b> | <b>1.807</b>  | <b>2.754</b>  |
| <b>HORARIA</b>  | <b>BO_SECO</b>  | <b>RF</b>  | <b>0.787</b> | <b>0.937</b> | <b>0.887</b> | <b>0.831</b> | <b>0.838</b>  | <b>0.787</b> | <b>1.795</b>  | <b>2.728</b>  |
|                 | <b>_HORARIA</b> |            |              |              |              |              |               |              |               |               |
|                 |                 | <b>GBR</b> | <b>0.783</b> | <b>0.936</b> | <b>0.885</b> | <b>0.829</b> | <b>0.838</b>  | <b>0.783</b> | <b>1.810</b>  | <b>2.756</b>  |
|                 |                 |            |              |              |              |              |               |              |               |               |
|                 |                 | <b>ANN</b> | <b>0.786</b> | <b>0.936</b> | <b>0.887</b> | <b>0.830</b> | <b>0.834</b>  | <b>0.786</b> | <b>1.801</b>  | <b>2.743</b>  |
| <b>UMIDADE_</b> | <b>TEMPERA</b>  | <b>SVM</b> | <b>0.309</b> | <b>0.672</b> | <b>0.682</b> | <b>0.459</b> | <b>0.390</b>  | <b>0.465</b> | <b>3.703</b>  | <b>6.725</b>  |
| <b>RELATIVA</b> | <b>TURA_DO</b>  |            |              |              |              |              |               |              |               |               |
| <b>_DO_AR_</b>  | <b>_AR_BUL</b>  | <b>DT</b>  | <b>0.467</b> | <b>0.786</b> | <b>0.684</b> | <b>0.538</b> | <b>0.540</b>  | <b>0.468</b> | <b>3.252</b>  | <b>5.769</b>  |
| <b>HORARIA</b>  | <b>BO_SECO</b>  | <b>RF</b>  | <b>0.470</b> | <b>0.786</b> | <b>0.686</b> | <b>0.540</b> | <b>0.539</b>  | <b>0.471</b> | <b>3.244</b>  | <b>5.752</b>  |
|                 | <b>_HORARIA</b> |            |              |              |              |              |               |              |               |               |
|                 | <b>_3d</b>      | <b>GBR</b> | <b>0.463</b> | <b>0.783</b> | <b>0.681</b> | <b>0.533</b> | <b>0.537</b>  | <b>0.464</b> | <b>3.265</b>  | <b>5.803</b>  |



|          |          |     |        |       |       |       |        |       |        |        |
|----------|----------|-----|--------|-------|-------|-------|--------|-------|--------|--------|
| UMIDADE_ | UMIDADE_ | SVM | 0.346  | 0.708 | 0.687 | 0.487 | 0.448  | 0.472 | 14.328 | 17.180 |
| RELATIVA | RELATIVA | DT  | 0.472  | 0.799 | 0.687 | 0.549 | 0.560  | 0.473 | 12.881 | 15.341 |
| _DO_AR_  | _DO_AR_  | RF  | 0.475  | 0.801 | 0.690 | 0.553 | 0.564  | 0.476 | 12.844 | 15.286 |
| HORARIA  | HORARIA_ | GBR | 0.470  | 0.801 | 0.687 | 0.550 | 0.567  | 0.471 | 12.895 | 15.349 |
|          | 7d       | ANN | 0.469  | 0.797 | 0.686 | 0.547 | 0.557  | 0.470 | 12.909 | 15.383 |
| UMIDADE_ | VENTO_DI | SVM | 0.346  | 0.708 | 0.683 | 0.483 | 0.444  | 0.466 | 14.392 | 17.263 |
| RELATIVA | RECAO_H  | DT  | 0.463  | 0.795 | 0.681 | 0.541 | 0.553  | 0.463 | 13.044 | 15.548 |
| _DO_AR_  | ORARIA_7 | RF  | 0.467  | 0.798 | 0.684 | 0.546 | 0.558  | 0.468 | 12.993 | 15.473 |
| HORARIA  | d        | GBR | 0.463  | 0.797 | 0.681 | 0.543 | 0.559  | 0.463 | 13.043 | 15.537 |
|          |          | ANN | 0.466  | 0.797 | 0.683 | 0.544 | 0.554  | 0.467 | 13.003 | 15.493 |
| VENTO_VE | PRESSAO  | SVM | 0.121  | 0.514 | 0.429 | 0.220 | 0.159  | 0.184 | 2.736  | 7.111  |
| LOCIDADE | _ATMOSF  | DT  | 0.181  | 0.535 | 0.425 | 0.228 | 0.185  | 0.181 | 2.641  | 6.906  |
| _HORARIA | ERICA_AO | RF  | 0.182  | 0.537 | 0.427 | 0.229 | 0.186  | 0.182 | 2.640  | 6.897  |
|          | _NIVEL_D | GBR | 0.175  | 0.531 | 0.418 | 0.222 | 0.177  | 0.175 | 2.651  | 6.954  |
|          | A_ESTACA | ANN | 0.181  | 0.538 | 0.426 | 0.229 | 0.186  | 0.181 | 2.641  | 6.900  |
|          | O_HORARI |     |        |       |       |       |        |       |        |        |
|          | A        |     |        |       |       |       |        |       |        |        |
| VENTO_VE | VENTO_VE | SVM | -0.368 | 0.462 | 0.281 | 0.130 | -0.065 | 0.079 | 1.582  | 7.064  |

|          |          |     |        |       |       |       |        |       |       |       |
|----------|----------|-----|--------|-------|-------|-------|--------|-------|-------|-------|
| LOCIDADE | LOCIDADE | DT  | 0.181  | 0.534 | 0.426 | 0.228 | 0.177  | 0.182 | 1.224 | 5.496 |
| _HORARIA | _HORARIA | RF  | 0.182  | 0.551 | 0.429 | 0.236 | 0.197  | 0.184 | 1.224 | 5.445 |
|          | _3d      | GBR | 0.167  | 0.553 | 0.415 | 0.229 | 0.205  | 0.172 | 1.234 | 5.494 |
|          |          | ANN | 0.184  | 0.523 | 0.431 | 0.225 | 0.170  | 0.186 | 1.222 | 5.502 |
| VENTO_VE | VENTO_VE | SVM | -0.067 | 0.437 | 0.336 | 0.147 | -0.001 | 0.113 | 2.536 | 7.570 |
| LOCIDADE | LOCIDADE | DT  | 0.114  | 0.400 | 0.342 | 0.137 | 0.030  | 0.117 | 2.310 | 7.171 |
| _HORARIA | _HORARIA | RF  | 0.114  | 0.399 | 0.342 | 0.137 | 0.030  | 0.117 | 2.310 | 7.171 |
|          | _7d      | GBR | 0.107  | 0.382 | 0.331 | 0.127 | 0.015  | 0.110 | 2.320 | 7.249 |
|          |          | ANN | 0.115  | 0.399 | 0.345 | 0.137 | 0.031  | 0.119 | 2.309 | 7.162 |

---

JEAN FIRMINO CARDOSO

**ESTIMATIVA FUTURA DE VARIÁVEIS HIDROLÓGICAS UTILIZANDO TÉCNICAS  
DE APRENDIZAGEM DE MÁQUINA: UMA ABORDAGEM COMPARATIVA**

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Engenharia Civil do Campus Agreste da Universidade Federal de Pernambuco – UFPE, na modalidade de artigo científico, como requisito parcial para obtenção do grau de Bacharel em Engenharia Civil.

**Área de concentração:** Hidrologia

Aprovado em 12 de Março de 2024.

**BANCA EXAMINADORA**

---

Prof. Dr. Saulo de Tarso Marques Bezerra (Orientador)  
Universidade Federal de Pernambuco

---

Prof. Dr. Gilson Lima da Silva (Avaliador)  
Universidade Federal de Pernambuco

---

Profa. M.Sc. Andreia Azevedo Abrantes de Oliveira (Avaliadora)  
Centro Universitário Tabosa de Almeida