

**UNIVERSIDADE DO VALE DO RIO DOS SINOS
UNIDADE ACADÊMICA DE GRADUAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO**

FABIANE KUHN

**ESTIMATIVA DO ÍNDICE DE EVAPOTRANSPIRAÇÃO COM BASE EM TÉCNICAS
DE APRENDIZADO DE MÁQUINA**

São Leopoldo
2021

FABIANE KUHN

**ESTIMATIVA DO ÍNDICE DE EVAPOTRANSPIRAÇÃO COM BASE EM TÉCNICAS
DE APRENDIZADO DE MÁQUINA**

Artigo apresentado como requisito parcial para
obtenção do título de Bacharel em Ciência da
Computação, pelo Curso de Ciência da Compu-
tação da Universidade do Vale do Rio dos Sinos
(UNISINOS)

Orientador(a): Prof. Dr. Gabriel de Oliveira Ra-
mos

Co-orientador: Profa. Dra. Regina Célia Espi-
nosa Modolo

São Leopoldo
2021

ESTIMATIVA DO ÍNDICE DE EVAPOTRANSPIRAÇÃO COM BASE EM TÉCNICAS DE APRENDIZADO DE MÁQUINA

Fabiane Kuhn¹ Gabriel de Oliveira Ramos²

Regina Célia Espinosa Modolo³

Resumo: O crescimento populacional, alinhado às mudanças climáticas e às crises hídricas, tem despertado a atenção para a agropecuária. O setor é responsável pelo uso de 70% da água mundial e de um desperdício de, aproximadamente, 50% desse total em processos de irrigação. Diversos métodos tecnológicos estão sendo desenvolvidos para minimizar esse impacto e colaborar com os Objetivos de Desenvolvimento Sustentável da ONU, dentro deles, a agricultura sustentável. Visando a otimização do uso de recursos hídricos, torna-se necessária a análise de evapotranspiração, visto que é a variável mais ativa do ciclo hidrológico e a principal componente no balanço hídrico dos ecossistemas agrícolas. Partindo da hipótese de que análises com aprendizado de máquina podem ser aplicadas para determinação de evapotranspiração e auxílio na tomada de decisão na irrigação, permitindo estimativas assertivas e sem dependência de uma grande variedade de dados, este artigo apresenta a aplicação das técnicas de Decision Tree Regressor, Random Forest, Artificial Neural Network e XGBoost para esse fim. Através de um dataset do Instituto Nacional de Meteorologia (INMET), os modelos foram treinados com base em equações amplamente validadas para cálculos de evapotranspiração. Após a rotina de testes, foi possível obter resultados satisfatórios, com MAE inferior a 0,0015, demonstrando a efetividade das técnicas computacionais para a estimativa de evapotranspiração.

Palavras-chave: Evapotranspiração. Irrigação. Aprendizado de máquina. Agricultura.

Abstract: In line with climate change and water crises, the population growth has drawn attention to agriculture. The sector is responsible for the use of 70% of the world's water and a waste of approximately 50% of this total in irrigation processes. Several technological methods are being developed to minimize this impact and collaborate with the UN Sustainable Development Goals, within them, sustainable agriculture. Aiming at optimizing the use of water resources, it is necessary to analyze evapotranspiration, as it is the most active variable in the hydrological cycle and the main component in the water balance of agricultural ecosystems. Based on the hypothesis that Machine Learning analysis can be applied to determine evapotranspiration and aid in decision making in irrigation, allowing assertive estimates and without dependence on a wide variety of data, this article presents the application of Decision Tree Regressor techniques, Random Forest, Artificial Neural Network and XGBoost for that purpose. Using a dataset from the National Institute of Meteorology (INMET), the models were trained based on widely validated equations for Evapotranspiration calculations. After the testing routine, it was possible to obtain satisfactory results, with MAE less than 0.0015, demonstrating the effectiveness of computational techniques for estimating evapotranspiration.

Keywords: Evapotranspiration. Machine Learning. Irrigation. Agriculture.

¹Graduanda em de Ciência da Computação pela Unisinos. Email: fabianekuhn@edu.unisinos.br

²Orientador. Professor do Programa de Pós-Graduação em Computação Aplicada da Unisinos. Email: gdoramos@unisinos.br

³Co-orientadora. Professora e coordenadora do curso de Engenharia Agrônômica na Unisinos. Email: regi-naem@unisinos.br

1 INTRODUÇÃO

De acordo com a Organização Mundial da Saúde (OMS), a disponibilidade de água potável é fundamental para a saúde da população, seja ela utilizada para beber, lavar ou produzir alimentos. A preservação dos recursos hídricos deve ser uma prioridade na vida das pessoas (WHO, 2019). Contudo, atualmente 70% da água mundial é destinada à agricultura e, aproximadamente 50% desse total, é desperdiçado em processos de irrigação devido à falta de tecnologia e informações para os produtores (FAO, 2015).

A água é um dos principais fatores a ser controlado para a garantia de boas safras agrícolas, visto que tanto o excesso quanto a falta do recurso podem prejudicar a produtividade. Nesse contexto, torna-se necessária a adoção de tecnologias no agronegócio para o apoio na tomada de decisão. A agricultura 4.0 é uma abordagem que permite a utilização de dados para a análise de cenários, incorporando aspectos desenvolvidos para a indústria 4.0 no aspecto rural (RIBEIRO; MARINHO; ESPINOSA, 2017). Dentro desse contexto, a tecnologia pode ser utilizada para que apenas a quantidade necessária de água seja aplicada em campo, contribuindo de forma positiva para a economia e para a sustentabilidade da agricultura.

Os 17 Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas (ONU) integram, de forma equilibrada, três dimensões de desenvolvimento: social, econômica e ambiental (ONU, 2015). A agenda de ações visa construir um futuro melhor e a agricultura faz parte dessa meta. O ODS número dois possui o objetivo de agricultura sustentável e fome zero, fortemente relacionado à aplicação de tecnologia para o domínio de dados e decisões assertivas:

Até 2030, garantir sistemas sustentáveis de produção de alimentos e implementar práticas agrícolas robustas, que aumentem a produtividade e a produção, que ajudem a manter os ecossistemas, que fortaleçam a capacidade de adaptação às mudanças do clima, às condições meteorológicas extremas, secas, inundações e outros desastres, e que melhorem progressivamente a qualidade da terra e do solo. (ONU, 2015).

Além dos aspectos ambientais, a agricultura possui grande relevância econômica em todo o mundo. Segundo dados e cálculos do Centro de Estudos Avançados em Economia Aplicada (CEPEA), em parceria com a Confederação da Agricultura e Pecuária do Brasil (CNA), em 2020 o agronegócio brasileiro obteve a marca de 26,6% do Produto Interno Bruto (PIB) do Brasil. Ao comparar o desempenho do agronegócio em períodos anteriores, como 1970 (7,5%) e 2019 (20,5%), é possível observar o avanço do setor. (CEPEA, 2020). Dentro desse total, 68% de participação é da agricultura e 32% da pecuária.

Para que a produção seja sustentável e eficiente é necessário atuar em melhorias no processo de irrigação, responsável pela maior parte do uso da água e fundamental para a determinação de produtividade nas safras. O manejo correto da irrigação, através da tecnologia, pode ser baseado em dados coletados na lavoura e no uso de sistemas computacionais. A evapotranspi-

ração (ET), soma da evaporação da água pela superfície de solo com a transpiração da planta, é um fator decisivo para a tomada de decisão, sendo considerada a variável mais ativa do ciclo hidrológico e a principal componente no balanço hídrico dos ecossistemas agrícolas (PEREIRA VILLA NOVA, 1997).

Atualmente existem formas diretas e indiretas de realizar a determinação de evapotranspiração de referência, base para a análise de ET em conjunto com o coeficiente da cultura (K_c). A forma direta, através de equipamentos em campo, como lisímetros, possui custo elevado e um complexo manuseio, além de limitações técnicas, não apresentando informações precisas em situações de muita chuva (SANTIAGO, 2001). Já os métodos indiretos, formados por equações matemáticas, dependem de uma alta variedade de dados, por vezes indisponíveis aos produtores (ANTONOPOULOS; ANTONOPOULOS, 2017). A partir da necessidade de melhorias no custo-benefício, técnicas computacionais indiretas começam a ganhar destaque, como a aplicação de aprendizado de máquina.

Dentro desse contexto, há um número crescente de trabalhos relacionados com a aplicação de aprendizado de máquina para a análise de evapotranspiração. A pesquisa realizada por Yusuf, Abidakun e Makinde (2020) implementa o modelo de Multilayer Perceptron Neural Network (MLP-NN) para trabalhar com mais de 35 anos de dados, obtendo erro quadrático médio de 0,0031. Contudo, em seu dataset, o trabalho conta com informações de chuva, que por vezes não estão disponíveis em estações climáticas públicas, ou, quando o dado está disponível, não representa uma informação confiável para a região coberta pelo equipamento devido à má distribuição da precipitação, o que restringe as possibilidades de aplicação.

Sendo assim, este trabalho surge com intuito de entender como estimar a taxa de evapotranspiração, com boa precisão, a partir de um conjunto limitado de dados públicos e utilizando técnicas de aprendizado de máquina. Para tanto, o objetivo é implementar e comparar modelos de Artificial Neural Network, Random Forest, Decision Tree Regressor e XGBoost, com objetivo de aproximar os valores de métodos consagrados na literatura. A análise será embasada e validada através dos principais métodos indiretos para o cálculo de evapotranspiração horária, sendo eles Penman-Monteith, Makkink e Turc. As principais contribuições do trabalho envolvem os itens:

- Criação de pipeline para pré-processamento de dados e treinamento dos modelos;
- Comparação de resultados computacionais com métodos matemáticos consolidados na literatura;
- Desenvolvimento de tese para democratização ao acesso da informação de evapotranspiração através de dados públicos.

Este trabalho está estruturado em 6 seções. A seção 2 é focada em apresentar os principais conceitos encontrados na revisão bibliográfica sobre os assuntos envolvidos no presente estudo, detalhando os temas: irrigação, evapotranspiração, métodos tradicionais para análise de evapotranspiração e aprendizado de máquina. Na seção 3 são apresentados os trabalhos relacionados

através de uma revisão sistemática da literatura com aplicação da metodologia de Petersen, Vakkalanka e Kuzniarz (2015). A seção 4 apresenta o detalhamento do modelo proposto, com o diagrama completo do funcionamento da implementação. Na seção 5, são apresentadas as métricas, a baseline e os resultados obtidos com a implementação dos modelos computacionais. Por fim, a seção 6 apresenta as conclusões e recomendação de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Com o intuito de adentrar nas explicações a respeito do modelo computacional proposto e seu impacto, torna-se necessário obter embasamento nos conceitos que o permeiam. Sendo assim, essa seção visa apresentar as definições a respeito dos tópicos de irrigação, evapotranspiração e aprendizado de máquina.

2.1 Irrigação e o Impacto da Evapotranspiração

A irrigação é o processo de adicionar água em campo para suprir as demandas hídricas da cultura quando não há presença suficiente de chuva. O manejo correto da irrigação é extremamente importante para que se tenha um bom resultado de produtividade e redução de gastos. O manejo ocorre através da determinação da frequência de irrigação, do volume de água através da lâmina, da velocidade e intensidade da aplicação (LUCAS, 2019).

A irrigação permite maior segurança da produção, além de indicadores melhores de produtividade e de gerar a possibilidade de plantio na entressafra. Segundo dados da Agência Nacional de Águas (ANA), lavouras irrigadas chegam a produzir em média 3 vezes mais que o sequeiro (ANA, 2017).

A estimativa de evapotranspiração é crucial para o processo de irrigação. Visando obter uma alta precisão na análise, deve-se considerar as diferenças de cada cultura e suas respectivas fases fenológicas, visto que há mudança no processo biológico da planta (DANIEL et al., 2020).

A evapotranspiração é amplamente utilizada na agricultura e representa a perda de água para a atmosfera, que ocorre através da soma da evaporação com a transpiração da planta (DANIEL et al., 2020). A evaporação é caracterizada pela passagem da água contida no solo em estado líquido para gasoso através de uma temperatura inferior ao ponto de ebulição. Já a transpiração representa a evaporação da água consumida, processada e eliminada pelas plantas em seus processos biológicos (RODRIGUES, 2017).

Para a análise completa do processo de evapotranspiração os seguintes conceitos são aplicados (RODRIGUES, 2017):

- **Evapotranspiração potencial (Etp):** Representa a quantidade de água que poderá evapotranspirar em condições nas quais a umidade disponível no solo continua sendo suficiente para suprir as demandas hídricas da planta.

- Evapotranspiração real (Etr): Representa a evapotranspiração que ocorreu de fato, considerando condições atmosféricas, conteúdo de água no solo e características da vegetação. Entende-se que a Etr nunca pode ser superior a Etp.
- Evapotranspiração de Referência (Et0): Possui lógica similar a Etp, porém, é aplicada para uma cultura já identificada e conhecida, a grama. Considera a evapotranspiração dela quando já está adaptada ao solo.
- Evapotranspiração da cultura (Etc): Representa a união da Et0 com o coeficiente cultural (Kc) de cada cultivo. Esse valor é utilizado na maioria das análises práticas.

A evapotranspiração pode ser afetada por aspectos relacionados ao tipo de solo, vegetação e clima (DANIEL et al., 2020). A medição pode ser realizada de forma direta através de dispositivos mecânicos robustos, como tanques de água classe A ou lisímetros, que serão explicados em detalhes no item 2.2.1 deste trabalho. Essas soluções possuem um alto custo e uma difícil manutenção, sendo indicadas apenas para pesquisas e experimentos (RODRIGUES, 2017). Em cenários de rotinas de campo, o mais comum é o uso de fórmulas para a estimativa de evapotranspiração através de cálculos relacionando condições climáticas e evapotranspiração potencial. As fórmulas mais conhecidas são Penman-Monteith, Hargreaves-Samani, Priestley-Taylor, Makkink e Turc, que possuem algumas variações na manipulação dos dados utilizados (ANTONOPOULOS; ANTONOPOULOS, 2017). A fórmula de Penman-Monteith foi definida como padrão pela Food and Agriculture Organization (FAO - ONU). Contudo, a abordagem demanda diversos dados climáticos de entrada, como saldo de radiação na superfície, fluxo de calor sensível no solo, temperatura média do ar a 2 metros de altura, velocidade do vento a 2 metros de altura, pressão de saturação de vapor e chuva, e possui um cálculo matematicamente complexo (RODRIGUES, 2017).

2.2 Métodos tradicionais para análise de ET

Os principais métodos utilizados para a análise de evapotranspiração podem ser divididos em duas categorias: medição *in situ* de ET0, através de lisímetros, e estimativas de ET0, através de equações.

2.2.1 Medição *in situ* de evapotranspiração

A principal forma de medição de evapotranspiração em campo é através de equipamentos chamados de lisímetros, que são grandes recipientes instalados em condições de campo e cheios de solo (MORAES; RENÊ; JÚNIOR, 2009). Os mesmos possuem um sistema de pesagem ou um sistema de controle de entrada e saída de água e podem apresentar uma superfície coberta com vegetação ou nua.

A Figura 1 representa a estrutura de um lisímetro de pesagem em campo. As versões mais recentes permitem que a variação de massa seja medida por células de carga (SANTIAGO, 2001). Essa abordagem faz que com as células variem sua resistência de acordo com a microdeformação que é causada pela variação de massa sustentada no equipamento. Assim obtém-se uma medida detalhada da variação de massa de água no ambiente controlado. Contudo, esse método possui um difícil manuseio. Além disso, em períodos de chuvas intensas e prolongadas os resultados de evapotranspiração são questionáveis (SANTIAGO, 2001). Mesmo em dias sem chuva, onde a medida se torna mais confiável, há a impossibilidade de obtenção de medidas em escalas curtas de tempo, como em uma escala horária.

Por esses motivos, os lisímetros são utilizados em ambientes controlados e de pesquisa, no qual há mão de obra extremamente qualificada para o manuseio e o propósito não é a tomada de decisão na rotina de irrigação em campo (SANTIAGO, 2001).

Figura 1 – Lisímetro de pesagem



Fonte: Estudo de Nascimento et al. (2011)

2.2.2 Estimativa de ET_0 através de equações

Devido à alta complexidade de utilização e custo elevado para estruturas de medição *in situ* de evapotranspiração, a abordagem mais comum é o uso de equações para a estimativa dos índices com base em informações climáticas. As próximas subseções irão apresentar três equações que possuem o propósito de determinar a evapotranspiração.

2.2.2.1 Penman-Monteith

A equação Penman-Monteith, também conhecida como FAO-56, é a alternativa estabelecida como padrão pela ONU para a Alimentação e a Agricultura (ANTONOPOULOS; ANTONOPOULOS, 2017) e é considerada a equação com maior precisão. A mesma possui uma alta

demanda de dados para seu cálculo, conforme equação 1. Além disso, para a obtenção dos valores através de estações meteorológicas comuns, torna-se necessário diversos cálculos complementares complexos, apresentados de forma detalhada no Apêndice A:

$$ET_0 = \frac{0,408 \times \Delta(R_n - G) + \gamma \frac{37}{T_{med} + 237} \times u_2(e_s - e_a)}{\Delta + \gamma(1 + 0,34 \times u_2)} \quad (1)$$

Onde:

ET_0 = Evapotranspiração de referência com base horária, ou seja, evapotranspiração no período de uma hora (mm h^{-1})

Δ = Inclinação da curva de pressão de vapor ($\text{KPa } ^\circ\text{C}^{-1}$)

R_n = Saldo de radiação ($\text{MJ m}^2\text{h}^{-1}$)

G = Fluxo de calor no solo - Nesse caso, considera-se 0 ($\text{MJ m}^2\text{h}^{-1}$).

γ = Constante psicométrica ($\text{KPa } ^\circ\text{C}^{-1}$)

T_{med} = Temperatura média ($^\circ\text{C}$)

u_2 = Velocidade do vento a 2 metros de altura ($\text{m } 2^{-1}$)

e_s = Pressão de saturação do vapor d'água (KPa)

e_a = Pressão de vapor d'água (KPa)

Esse cálculo, mesmo que assertivo, possui uma elevada complexidade que, por vezes, não representa a realidade dos produtores rurais. Segundo dados do levantamento Censo Agro 2017, 70% dos produtores rurais possuem, no máximo, o ensino fundamental incompleto, dificultando o entendimento de procedimentos matemáticos complexos (IBGE, 2017).

2.2.2.2 Makkink

O método Makkink está enquadrado nos modelos baseados em radiação e utiliza a equação 2 (ZHU et al., 2020):

$$ET_0 = \frac{a}{\gamma} \frac{\Delta}{\Delta + \gamma} (R_s - b) \quad (2)$$

Onde:

ET_0 = Evapotranspiração de referência com base horária, ou seja, evapotranspiração no período de uma hora (mm h^{-1})

a, b = Constantes empíricas

Δ = Inclinação da função de pressão de vapor de saturação ($\text{KPa } ^\circ\text{C}^{-1}$)

γ = Calor latente de vaporização ($\text{MJ m}^2\text{h}^{-1}$)

R_s = Radiação solar ($\text{MJ m}^2\text{h}^{-1}$)

2.2.2.3 Turc

A equação de Turc varia de acordo com a umidade relativa (KAYA; MAMAK; UNES, 2016), quando a umidade relativa é menor que 50%, é utilizada a fórmula 3:

$$ET0 = 0.0133 \frac{T_a}{T_a + 15} (R_s + 50) \quad (3)$$

Quando a umidade relativa é maior que 50%, a fórmula passa a ser a 4:

$$ET0 = 0.0133 \frac{T_a}{T_a + 15} (R_s + 50) \left(1 + \frac{50 - U_r}{70}\right) \quad (4)$$

Onde:

$ET0$ = Evapotranspiração de referência com base horária, ou seja, evapotranspiração no período de uma hora (mm h^{-1})

T_a = Temperatura média ($^{\circ}\text{C}$)

R_s = Radiação solar ($\text{MJ m}^2\text{h}^{-1}$)

U_r = Umidade relativa (%)

2.2.3 Discussão sobre os métodos

A aplicação de equações para a análise de evapotranspiração permite a obtenção de resultados aceitáveis com um baixo custo. Contudo, esses métodos demandam uma grande variedade de dados, que por vezes não estão disponíveis de forma gratuita para os produtores rurais. Além da dificuldade de acesso aos dados, as equações possuem um nível de complexidade matemática que não corresponde à média de escolaridade dos agricultores brasileiros. Esse cenário demanda a aplicação de novas abordagens, que permitam a análise de evapotranspiração de forma simples e com dados disponíveis publicamente, gerando maior apoio prático à tomada de decisão na irrigação.

2.3 Aprendizado de Máquina

O aprendizado de máquina, ou Machine Learning, é uma divisão da Inteligência Artificial que representa o processo de aprendizado contínuo de uma máquina para resolução de problemas (SILVA, 2017). O processo funciona a partir do fornecimento de dados de entrada para que o sistema aprenda através de rotinas de treinamento e passe a gerar capacidade de apresentar saídas satisfatórias para o cenário (DAMACENO; VASCONCELOS, 2018). Os algoritmos são estruturados com equações iniciais pré-definidas e funcionam de forma semelhante ao funcionamento do aprendizado humano, quanto mais dados forem fornecidos para a máquina, mais

assertiva será sua saída, reduzindo as taxas de erros através do aprimoramento do conhecimento. As técnicas de aprendizado podem ser consideradas supervisionadas ou não supervisionadas. No aprendizado supervisionado o modelo aprende com base em resultados pré-definidos, utilizados como referência de treinamento para a aplicação. Já na abordagem não supervisionada, o modelo não possui resultados pré-definidos como referência no aprendizado (SILVA, 2017).

Existem diferentes métodos de aprendizado de máquina e, para este trabalho, serão utilizados quatro modelos: Decision Tree Regressor, Random Forest, Artificial Neural Network e XGBoost.

2.3.1 Decision Tree Regressor

A árvore de decisão está enquadrada dentro da classe de aprendizagem supervisionada e é construída através da divisão dos dados de origem em subconjuntos. O processo de divisão de dados é realizado repetidas vezes através de recursão. A recursão é considerada concluída quando o subconjunto de um nó tem os mesmos valores da variável de destino, ou ainda quando a divisão não agrega mais valor ao processo (ELAVARASAN et al., 2018). Esse modelo é extremamente popular e tem como vantagem a sua facilidade de implementação e de entendimento.

2.3.2 Random Forest

O modelo *Random Forest*, ou Floresta Aleatória, é classificado como *Ensemble*, ou seja, é um algoritmo de aprendizagem baseado em conjunto. Esse modelo utiliza diversas árvores de decisão como base e faz a previsão calculando a média das previsões de cada uma das árvores que a compõem. Sendo assim, muitas vezes possui um resultado melhor que apenas uma árvore de decisão (ELAVARASAN et al., 2018).

2.3.3 Artificial Neural Network

ANNs são modelos baseados no cérebro humano, projetados para simular a forma como as informações são interpretadas e processadas, através da detecção de padrões e do relacionamento entre dados (AGATONOVIC-KUSTRIN; BERESFORD, 2000). O modelo da ANN é formado a partir de centenas de unidades únicas, que são chamadas de neurônios, conectadas com pesos que criam a estrutura neural. Cada neurônio pode realizar funções simples de processamento, contudo, o verdadeiro poder está na conexão deles em rede (AGATONOVIC-KUSTRIN; BERESFORD, 2000).

2.3.4 XGBoost

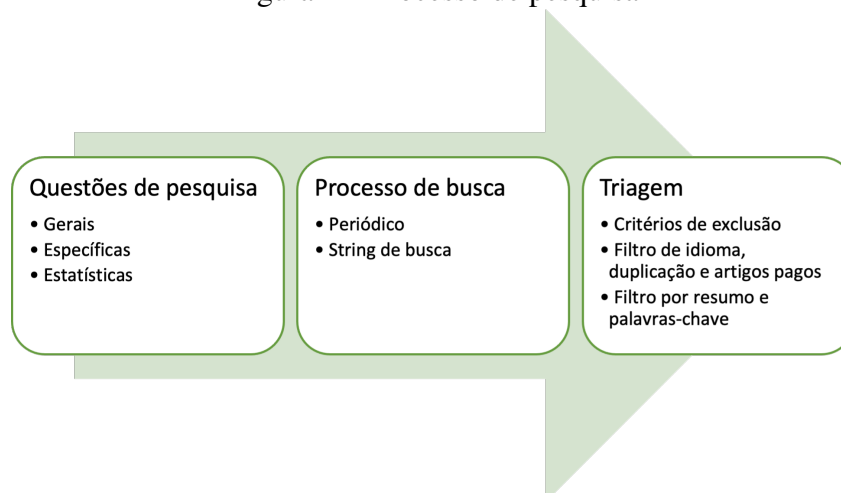
XGBoost significa Extreme Gradient Boosting, e se enquadra como um algoritmo baseado em árvores de decisão com aumento de gradiente. O aumento de gradiente permite a minimização das perdas enquanto novos modelos vão sendo adicionados. Isso ocorre porque o método passa por ciclos para adicionar modelos de forma interativa em um conjunto. Para isso, em um primeiro momento é iniciado o conjunto com um único modelo e com previsões básicas, então se inicia o ciclo onde é feita a previsão com esse modelo, calculada a função de perda, realizado o ajuste no modelo através dessa perda e se repete o ciclo. No modelo XGBoost há a aplicação de recursos adicionais com foco em velocidade e desempenho. A aplicação de XGBoost está ganhando muito espaço pelos seus resultados (BROWNLEE, 2016).

3 TRABALHOS RELACIONADOS

Esta seção visa apresentar trabalhos desenvolvidos e publicados anteriormente que se relacionam com o problema da presente pesquisa. Para uma maior amostragem e assertividade, a análise de trabalhos relacionados foi realizada através de uma Revisão Sistemática da Literatura relacionada à aplicação de técnicas de aprendizado de máquina para análise de evapotranspiração e apoio na tomada de decisão na irrigação. A estrutura da revisão foi embasada na pergunta norteadora "Quais são os métodos de aprendizado de máquina utilizados para determinação de evapotranspiração?"

Para a revisão sistemática da literatura foi implementada a metodologia descrita em Petersen, Vakkalanka e Kuzniarz (2015), que apresenta o processo de mapeamento sistemático através do levantamento de questões de pesquisa, do processo de busca e da triagem de resultados. A Figura 2 descreve as etapas de pesquisa implementadas.

Figura 2 – Processo de pesquisa



Fonte: Elaborado pela autora, adaptado de Petersen, Vakkalanka e Kuzniarz (2015).

3.1 Questões de Pesquisa

As questões de pesquisa foram organizadas em três grupos: Questão geral, questões específicas e questões estatísticas, conforme Quadro 1.

Quadro 1 – Questões de pesquisa	
Questão geral	
G1	Quais são os métodos de ML para determinação de ET?
Questões específicas	
E1	Quais features são utilizadas?
E2	Quais os métodos consolidados utilizados como baseline?
E3	Qual o período de dados utilizados nas features?
E4	Quais os principais datasets utilizados?
E5	Quais métricas são utilizadas para validação?
E6	Quais os índices de erro encontrados?
Questão estatística	
S1	Quantas publicações foram realizadas por ano?

Fonte: Elaborada pela autora.

3.2 Processo de Busca

A busca foi realizada na base de dados Google Scholar em abril de 2021, seguindo os termos do Quadro 2. A escolha da base se justifica no acesso direto aos artigos indexados em diferentes bases e fontes de conhecimento, contudo, devido as limitações do Google Scholar por permitir apresentação de artigos e materiais sem revisão por pares, os filtros aplicados na metodologia de análise foram fundamentais para garantir a relevância da revisão da literatura.

Termo	Sinônimos
Evapotranspiration	“Evapotranspiration”
Irrigation	“irrigation” OR “sprinkling” OR "watering"
Machine Learning	“machine learning” OR "deep learning"OR "neural network"OR “supervised learning” OR “artificial intelligence”

Fonte: Elaborada pela autora.

Através desses termos a String de busca utilizada foi: ((“Evapotranspiration”) AND (“irrigation” OR “sprinkling” OR "watering") AND (“machine learning” OR "deep learning"OR "neural network"OR “supervised learning” OR “artificial intelligence”)).

3.3 Triagem

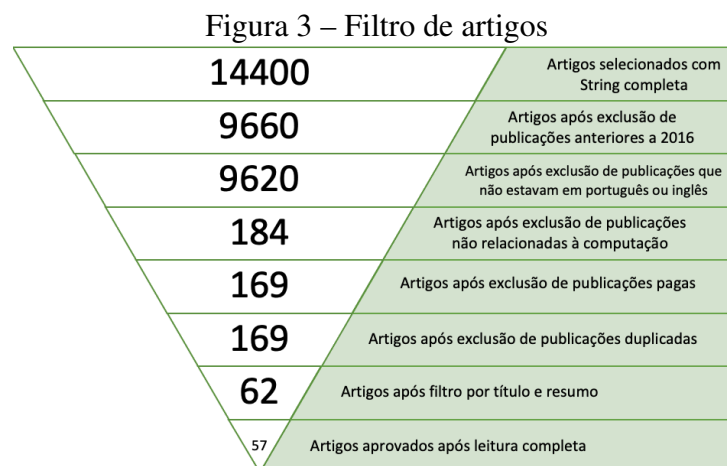
Para obtenção de maior assertividade na busca, foram utilizados três critérios de exclusão, sendo eles:

- Publicação anterior à 2016, visando trabalhar com modelos atualizados de aplicações;
- Estudos publicados em categorias não relacionadas à computação dentro dos periódicos, como por exemplo: hidrologia, agronomia e ciências do solo, por pesquisadores que não possuem relação com a área da computação. A análise dos pesquisadores foi realizada pela observação de área de atuação no artigo.
- Estudos sem resultados práticos a nível computacional.

Considerando a String de busca e os critérios de exclusão, 169 artigos foram selecionados. Após essa etapa, foram submetidos à novos filtros e análises:

- Eliminação de artigos disponibilizados em idiomas diferentes de português e inglês;
- Eliminação de artigos pagos;
- Eliminação de artigos duplicados;
- Realização de análise por título e resumo buscando pela relação de aprendizado de máquina para análise de evapotranspiração;
- Realização de leitura completa do artigo para entendimento da relação de aprendizado de máquina para evapotranspiração.

A Figura 3 representa o comportamento dos filtros de artigos e a obtenção de 57 artigos para análise final.



Fonte: Elaborado pela autora.

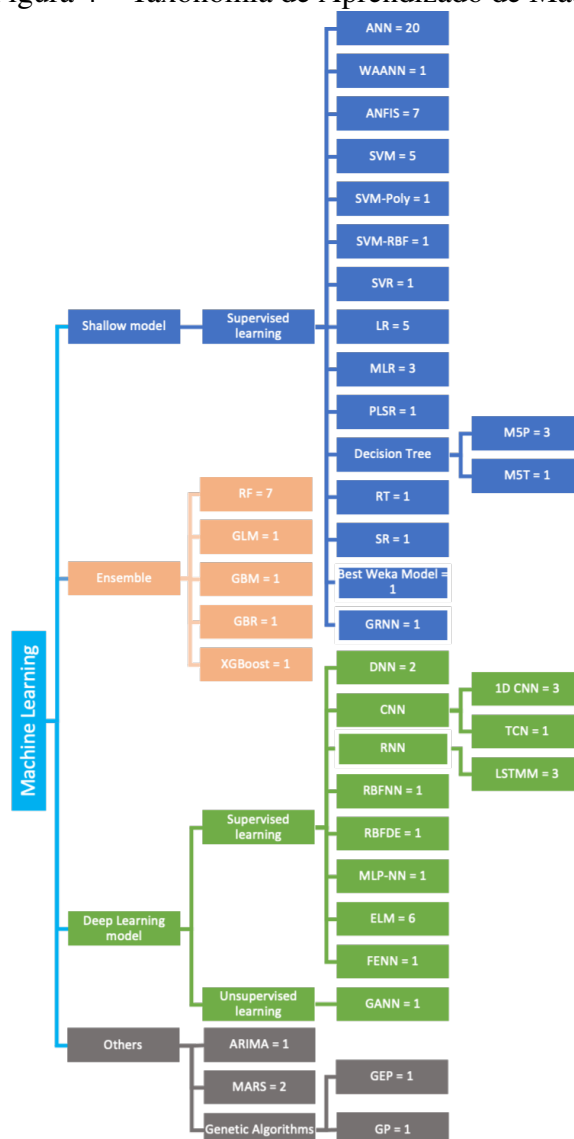
3.4 Análise dos artigos selecionados

Após a aplicação de filtros e análises para a seleção dos artigos, os materiais foram avaliados com base nas questões de pesquisa, disponíveis no Quadro 1. O objetivo da revisão sistemática da literatura, que contou com 57 artigos selecionados, é encontrar os padrões utilizados nos estudos.

3.4.1 Modelos de Aprendizado de Máquina

Os artigos analisados apresentaram um ou mais modelos de aprendizado de máquina para a análise de evapotranspiração, que podem ser observados na taxonomia desenvolvida na Figura 4, assim como o número de ocorrências dentro dos artigos.

Figura 4 – Taxonomia de Aprendizado de Máquina



Fonte: Elaborado pela autora.

É possível observar uma grande variedade de modelos empregados, em sua grande maioria com abordagem de aprendizado supervisionado, onde há a utilização de padrões pré-definidos para o treinamento dos modelos. O fato do ambiente científico contar com um grande número de estudos validando modelos de aprendizado de máquina diferentes para a análise de evapotranspiração evidencia que ainda há espaço para novas descobertas sobre qual o estado da arte nesse cenário.

3.4.2 Features

Os 57 estudos analisados contaram com diferentes combinações de features para alimentação dos processos de aprendizado de máquina, trabalhando tanto com dados climáticos quanto com informações de solo e imagem. A principal feature utilizada foi Temperatura, presente em 52 estudos, seguida de Velocidade do vento com 46 aparições e de Umidade relativa do ar com 43 usos. As ocorrências podem ser observadas no Quadro 3.

Quadro 3 – Features utilizadas

Feature	Estudos
Temperatura	52
Velocidade do vento	46
Umidade relativa do ar	43
Radiação solar	29
Chuva	16
Horas de sol	14
Pressão atmosférica	6
Umidade do solo	4
Déficit de pressão de vapor	4
Temperatura do solo	4
Radiação líquida	4
Radiação global	3
Altura da planta	2
LAI - Índice de área foliar	2
Radiação extraterrestre	2
Features que aparecem em apenas um estudo	25

Fonte: Elaborada pela autora.

Conforme a Quadro 3, 25 features foram utilizadas uma única vez dentre os artigos selecionados, não representando grande interferência na análise de evapotranspiração. Essas features são: Dia do ano, Profundidade da água do pantanal, Água de irrigação ecológica, Água drenada e escoamento, Higrometria, evaporação da panela (tanque classe A), Altitude, Direção do vento, Visibilidade, SPEI - índice padronizado de evapotranspiração de precipitação, SOI, BEST, Nino, Radiação de onda curta, Cobertura de núvens, Fluxo de calor no solo, Cobertura do solo, Evaporação, Potencial de água no solo, Ponto de orvalho, NDVI, SAVI, VVI, Velocidade

da rajada e estágio de crescimento da planta.

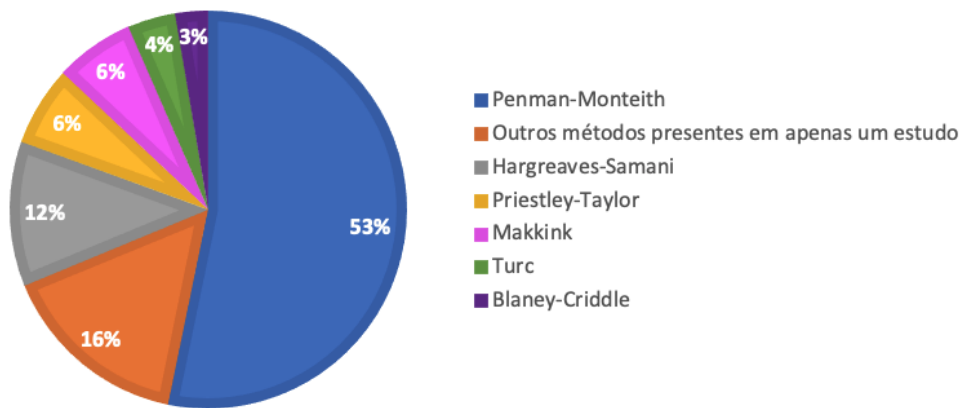
É possível perceber que as features mais utilizadas dentro dos estudos são aquelas que tem influência direta nas equações amplamente validadas e utilizadas como baseline: temperatura, velocidade do vento, umidade relativa do ar e radiação solar. Além do impacto direto, esses dados podem ser encontrados com facilidade em datasets públicos, disponíveis gratuitamente.

3.4.3 Métodos usados como baseline

As abordagens de aprendizado de máquina para a determinação de evapotranspiração contaram, em sua grande maioria, com comparações de métodos já validados para levantamento de índices de erro e de correlação. O principal método utilizado para comparação foi Penman-Monteith, isso porque é o modelo mais robusto e indicado oficialmente pela FAO (ANTONONOPOULOS; ANTONONOPOULOS, 2017). Essa abordagem esteve presente em 41 estudos. É necessário indicar que alguns artigos utilizaram mais de um método de comparação para a validação da proposta de aprendizado de máquina.

A distribuição de métodos padrões de comparação pode ser analisada na Figura 5.

Figura 5 – Análise de métodos de comparação



Fonte: Elaborado pela autora.

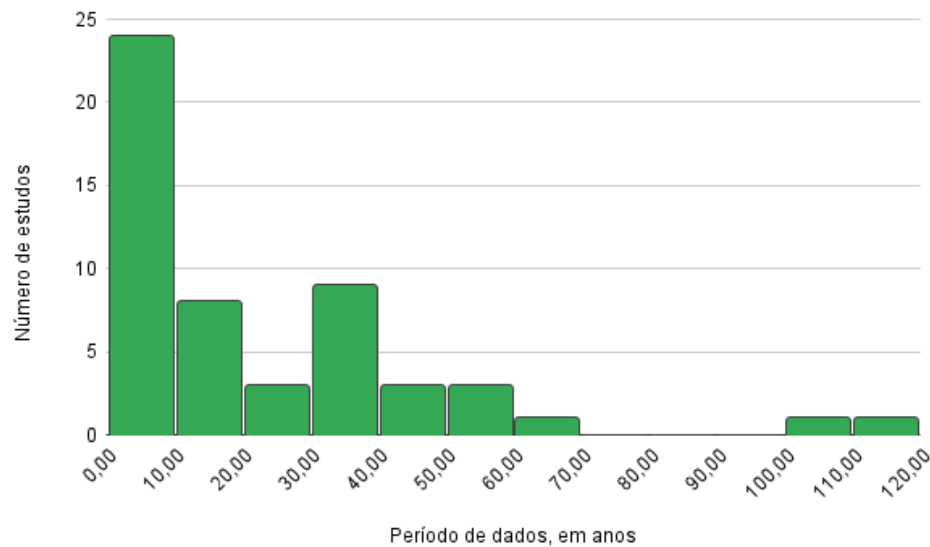
Em relação aos métodos utilizados como baseline, vale ressaltar a aplicação de cada um deles no que diz respeito a periodicidade de dados. Os modelos Blaney-Criddle e Hargreaves podem ser utilizados apenas para a análise de evapotranspiração diária, os demais métodos são aplicados tanto em base diária quanto horária (MINACAPILLI et al., 2016).

3.4.4 Período de dados

Todos os estudos analisados dentro da revisão sistemática da literatura contaram com features, ou seja, entrada de dados, de diferentes períodos e quantidades. O artigo com maior período

de dados analisados contou com 117 anos de informações, enquanto o artigo com o menor período teve apenas 3 meses de dados de entrada. A média entre todos artigos é de 21,16 anos. A distribuição de frequências do período de dados pode ser consultada no histograma apresentado na Figura 6.

Figura 6 – Histograma do período de dados



Fonte: Elaborado pela autora.

3.4.5 Datasets

Os estudos contaram com diferentes modelos de datasets, tanto de abordagens locais, como estações meteorológicas fixas em áreas experimentais, quanto de estações virtuais de cobertura nacional e global. De forma global, houve uma presença significativa do NASA Prediction of Worldwide Energy resource (POWER)⁴, utilizado em 4 estudos. A nível nacional, 5 artigos contaram com dados do Instituto Nacional de Meteorologia (INMET)⁵, que possui cobertura em todo o território brasileiro. O mesmo ocorreu com dados do India Meteorological Department⁶, apresentados em 5 trabalhos. Por fim, diversos trabalhos contaram com estações físicas privadas em um campo de testes, não trabalhando com dados públicos acessíveis para todos os produtores rurais.

⁴<https://power.larc.nasa.gov/data-access-viewer/>

⁵<https://mapas.inmet.gov.br/>

⁶<https://mausam.imd.gov.in>

3.4.6 Métricas

Os estudos avaliados nessa revisão da literatura contaram com métricas para análises comparativas entre o resultado obtido com a implementação de aprendizado de máquina em relação aos métodos padrões utilizados, como a equação de Penman-Monteith. O objetivo das métricas é determinar o erro obtido com modelos computacionais e também o índice de correlação com padrões calculados.

Dentre os 57 estudos, 40 contaram com a métrica Root Mean Square Error (RMSE) para validação. O cálculo é determinado pela equação 5:

$$RMSE = \sqrt{\frac{1}{n} \sum (P_i - O_i)^2} \quad (5)$$

A métrica do Coeficiente de Determinação foi utilizada por 34 estudos para a análise de correlação entre as estimativas computacionais para ET0 e os parâmetros calculados pelas equações, seguindo o padrão descrito na equação 6:

$$R^2 = \frac{(\sum (O_i - \bar{O}_i)(P_i - \bar{P}_i))^2}{\sum (O_i - \bar{O}_i)^2 \sum (P_i - \bar{P}_i)^2} \quad (6)$$

A terceira métrica com maior uso dentre os estudos é a Mean Absolute Error (MAE), presente em 30 trabalhos e representada pela equação 7:

$$MAE = \frac{1}{n} \sum |P_i - O_i| \quad (7)$$

Todas as equações possuem a representação onde:

- n = Número de dados
- P = Dado previsto
- O = Dado observado
- \bar{P} = Média do dado previsto
- \bar{O} = Média do dado observado

Através dessas e outras equações, os estudos analisados trabalham com métricas para determinar o sucesso da implementação de aprendizado de máquina para a análise de evapotranspiração e o nível de confiabilidade do processo. Todas as métricas mapeadas podem ser vistas no Quadro 4.

Conforme apresentado no Quadro 4, algumas métricas foram utilizadas em um único estudo, sendo elas: MAD, GPI, RMSD, E, SI, ACC, LL, c, RMSLE, a, TS, CE, Rm, b, BIAS, ME, RRMSE, SEP, DR, CC, IA. As métricas utilizadas com maior frequência são comuns em análises de aprendizado de máquina, tendo sua efetividade para a medição de erros amplamente validada.

Quadro 4 – Métricas utilizadas

Métrica	Estudos
RMSE	40
R2	34
MAE	30
r	13
NSE	7
MSE	6
NS	5
MBE	4
MAPE	4
SEE	3
d	3
RRMSE	3
MRE	2
MARE	2
nRMSE	2
Métricas que aparecem em apenas um estudo	21

Fonte: Elaborada pela autora.

3.4.7 Resultados obtidos

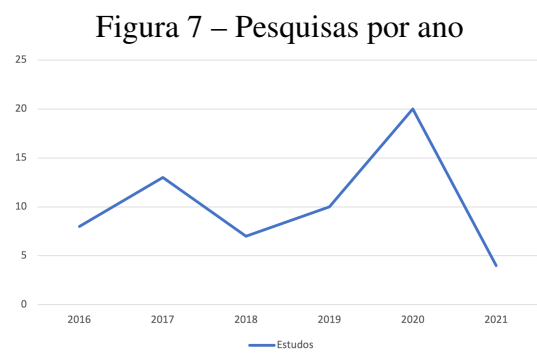
Ao analisar os resultados dos 57 artigos, é possível destacar dois casos extremos em relação ao erro encontrado:

- RMSE = 0,003162. O menor erro ocorreu no artigo de Yusuf, Abidakun e Makinde (2020) ao aplicar Multilayer Perceptron Neural Network (MLP-NN) com arquitetura 5-9-1, ou seja, 5 neurônios de entrada, 9 camadas intermediárias escondidas e 1 camada de saída. A aplicação contou com 35 anos de dados e o dataset empregado foi o NASA Prediction of Worldwide Energy resource (POWER). As features consideradas foram: Temperatura, Umidade Relativa, Radiação Extraterrestre, Chuva, Radiação solar e Velocidade do vento.
- RMSE = 0,88. O maior erro ocorreu no artigo de Ferreira e Cunha (2020) ao aplicar uma abordagem de CNN-LSTM, uma combinação de 1D CNN com LSTM. Foram utilizados 12 anos de dados com o dataset do Instituto Nacional de Meteorologia (INMET). Foram consideradas as seguintes features: Temperatura, Umidade relativa, Radiação solar e Velocidade do vento.

É possível perceber que as features utilizadas em ambos os estudos são semelhantes, não representando uma relação direta com a diferença de desempenho. O melhor cenário, com baixo erro, utilizou cerca de 3 vezes mais informações de período de dados que o pior caso. Com isso, conseguiu ampliar o treinamento e a validação das redes. Além disso, o método de aprendizado de máquina utilizado em cada caso é diferente.

3.4.8 Publicações por ano

Foram analisados 62 artigos após Strings de busca, critérios de exclusão e filtros, entrando na etapa de análise completa do conteúdo. De forma estatística, também foram analisados os anos de publicação desses estudos. A Figura 7 apresenta o gráfico de pesquisas selecionadas por ano.



Fonte: Elaborado pela autora.

É possível perceber um crescimento no volume de pesquisas nos últimos anos, representado na curva ascendente entre 2018 e 2020. Vale ressaltar que os dados de 2021 são parciais, uma vez que a busca por artigos foi realizada no mês de abril. Essa análise prova que as contribuições de aprendizado de máquina para a agricultura e, mais especificamente, para evapotranspiração têm crescido nos últimos anos.

3.5 Análise da revisão sistemática da literatura

Os métodos de aprendizado de máquina apresentam uma alternativa válida para a análise de evapotranspiração, conforme observado nos 57 estudos analisados. Diferentes métodos e abordagens vem sendo testados, em todo o mundo, alternando combinações de features e períodos de dados, e comparados com métodos tradicionais, ofertando alternativas coerentes aos métodos manuais.

Torna-se necessário destacar a grande influência do dado de Temperatura na análise de Evapotranspiração, devido sua direta relação com a evaporação da água do solo e transpiração da planta. Sendo assim, é compreendido o motivo da feature estar presente em 52 dos 57 estudos analisados e na maioria das equações tradicionais. Em casos de dados escassos, ou seja, baixa disponibilidade de features, alternativas de aprendizado de máquina ainda podem apresentar bons resultados, como é o caso do estudo de Lucas et al. (2018), que contou com apenas dados de Temperatura e dia do ano para a aplicação de ANN e obteve RMSE de 0,24 ao utilizar a equação de Hargreaves-Samani como método de comparação. Ainda no contexto de dados escassos, o estudo de Torres et al. (2020) utiliza apenas Umidade do solo e Temperatura para uma aplicação de SVM e possui RMSE = 0,79, um erro significativamente maior. Além do tipo

de método de aprendizado de máquina, os estudos divergem no período de dados aplicados. Enquanto a aplicação de ANN contou com 35 anos de dados, a SVM apenas com 4 meses. Essa comparação mostra que podem ser encontradas variações, tendo em vista a tomada de decisão de métodos, features e período de dados no momento da aplicação.

Com a análise dos artigos selecionados, percebe-se que ainda há espaço para a aplicação de novas combinações de métodos de aprendizado, features e datasets, visando maior assertividade em dados escassos. Dentro desse contexto, o presente trabalho possui o objetivo de utilizar dados públicos para a implementação de técnicas para análise de evapotranspiração.

4 MODELO PROPOSTO

O objetivo desta seção é descrever os parâmetros e arquitetura proposta para a aplicação de aprendizado de máquina para a análise de evapotranspiração.

4.1 Visão Geral

A estrutura lógica da aplicação está implementada de acordo com a Figura 8, onde são apresentadas as etapas de leitura do banco de dados, eliminação dos dados faltantes, eliminação dos dados inconsistentes, implementação da técnica de *scaling* para normalização dos dados no intervalo de 0 a 1, realização do processo de *split* entre dados de treinamento, com 70% da base, e dados de validação, com 30% da base, e por fim a implementação dos quatro modelos de aprendizado de máquina: Decision Tree Regressor, Random Forest Regressor, Artificial Neural Network e XGBoost. O código foi implementado na linguagem de programação Python e, assim como o banco de dados, está hospedado na plataforma Kaggle⁷.

4.2 Dataset

O dataset utilizado na aplicação foi obtido através do Banco de Dados Meteorológicos do INMET⁸ para a cidade de Luis Eduardo Magalhães - Bahia. A estação está instalada na localização 12°09'09.0"S 45°49'47.0"W, com altitude de 760,68 metros. A estação foi escolhida por estar presente em um dos maiores polos de irrigação do Brasil, gerando ainda mais relevância para as análises de evapotranspiração. O dataset possui medidas horárias no período de 22/04/2002 à 01/10/2020, com features relacionadas à temperatura, velocidade do vento, umidade relativa e radiação solar.

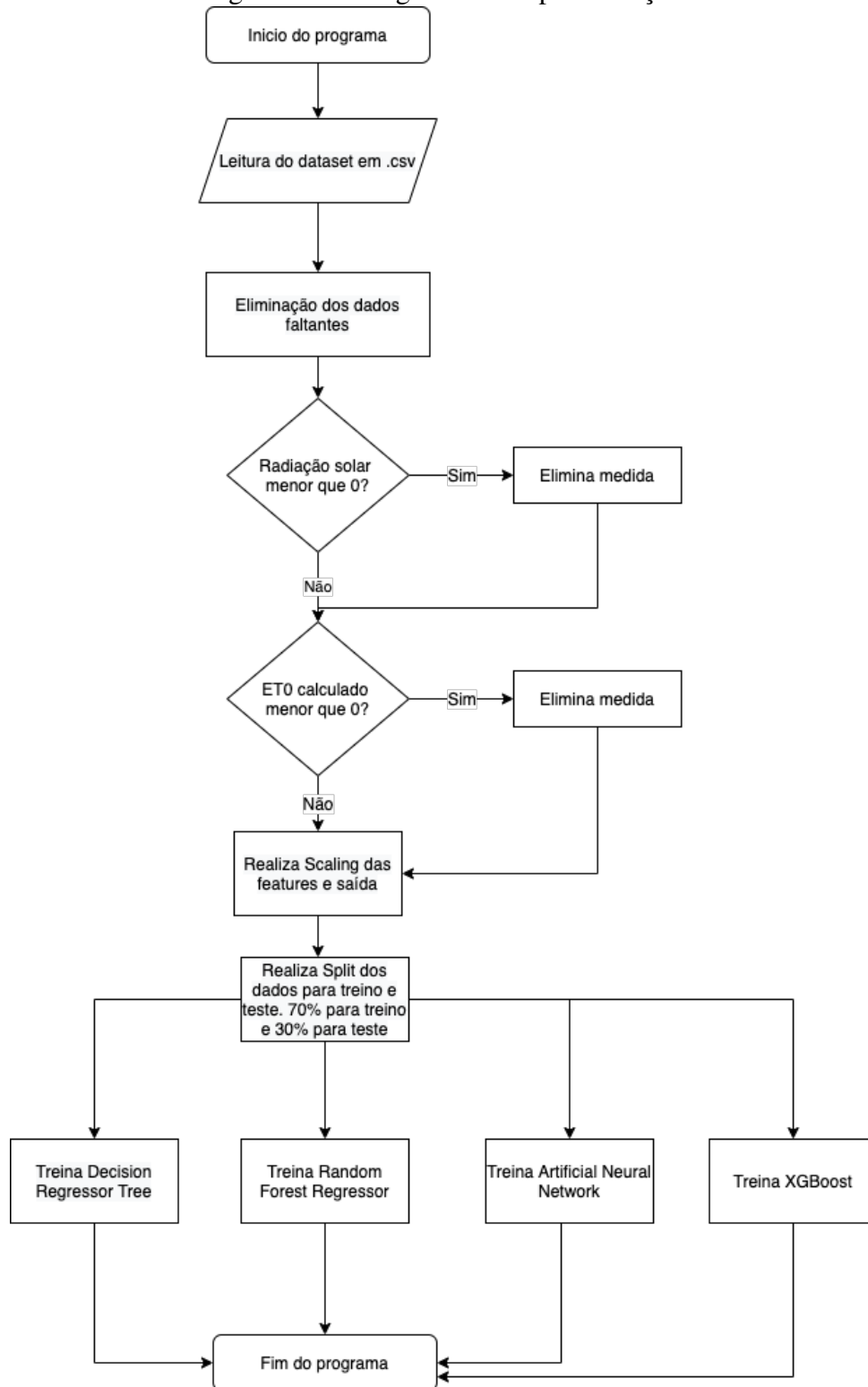
Para a geração dos dados de saída foram utilizados os três principais métodos indiretos para a análise de dados horários de evapotranspiração: Penman-Monteith (FAO-56), Makkink e Turc. Esses modelos podem ser aplicados tanto para análise diária quanto horária, diferentemente

⁷<http://kaggle.com>

⁸<https://bdmep.inmet.gov.br>

de equações como Hargreaves e Blaney–Criddle, que correlacionam apenas o resultado diário (MINACAPILLI et al., 2016).

Figura 8 – Fluxograma da implementação



Fonte: Elaborado pela autora.

4.3 Preparação dos dados

A preparação dos dados para o treinamento dos modelos de aprendizado de máquina ocorreu de acordo com a Figura 8, trabalhando com etapas de limpeza, organização e divisão de dados.

4.3.1 Leitura dos dados

A leitura dos dados ocorre pela interpretação de um arquivo .csv através da utilização da biblioteca *Pandas*, que possui foco na manipulação e análise de dados. A base de dados foi gerada através dos dados obtidos pelo INMET e adicionada na plataforma *Kaggle*. O dataset conta com 161.705 medidas.

4.3.2 Limpeza do dataset

Após a carga do dataset, torna-se necessário garantir que os valores existentes estão de fato coerentes. A base de dados oferecida pelo INMET é gerada através de uma estação meteorológica física instalada em campo, que coleta as informações de forma horária através de sensores. Sendo assim, existe propensão para falhas nas coletas e perdas nas medidas. Nesse cenário, é extremamente importante realizar a limpeza do dataset antes de realizar a implementação do processo de aprendizado de máquina.

Primeiramente, foi realizada a eliminação de medidas em que algum dos parâmetros fundamentais para os cálculos de evapotranspiração estivessem zerados. Esse processo foi realizado através da função *dropna* da biblioteca *Pandas*, que automaticamente realiza a análise da tabela e elimina linhas com dados nulos. Após esse processo, o dataset ficou com 146.225 medidas, tendo eliminado 15.480 linhas nessa etapa.

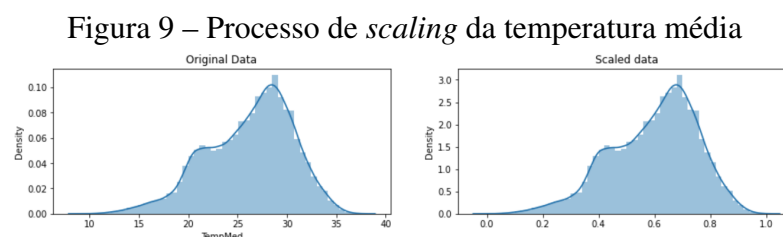
Após essa limpeza, foi implementada a eliminação de dados inconsistentes, que ocorrem quando há alguma falha nos sensores, mas ainda permitem que o dado seja enviado, com um valor que não reflete a realidade. Para isso, foi utilizada uma verificação lógica se dados como a Radiação solar estavam abaixo de zero, eliminando a ocorrência. Segundo Oliveira et al. (2013), os valores negativos podem ter como causa pequenos erros nos sensores piranômetros, responsáveis pela medida de radiação solar. Os erros ocorrem pelas mudanças graduais na atmosfera e radiação. Nesse momento o dataset chegou a 77.080 dados, tendo eliminado 69.145 nessa etapa e finalizando a limpeza. A grande quantidade de dados inconsistentes está alinhada à falhas temporárias no equipamento físico em campo.

4.3.3 Scaling dos dados

A aplicação da técnica de *min-max Scaling* dos dados tem como objetivo a realização do pré-processamento das informações para padronizar os intervalos de dados. Ao aplicar esse

processo a distribuição de dados continua a mesma, a única diferença está na escala, que no caso do presente trabalho é de 0 a 1. Essa padronização é importante, visto que algoritmos de aprendizado de máquina tendem a atribuir pesos maiores para valores maiores e pesos menores para valores menores, sem considerar as unidades de medida de cada feature (GUPTA, 2021).

A Figura 9 apresenta o resultado do processo de *scaling* de uma das features do dataset, a temperatura média.



Fonte: Elaborado pela autora.

4.3.4 Organização da base de treino e teste

Após a conclusão do preparo dos dados, os mesmos foram divididos no grupo de treino e teste através da função *split* da biblioteca de código aberto *scikit-learn*. A distribuição de dados foi aleatória e seguiu a regra de 70% da base para treino e 30% para teste, conforme melhores práticas citadas em Gholamy, Kreinovich e Kosheleva (2018).

4.3.5 Modelos de aprendizado

Os modelos de aprendizado de máquina escolhidos para a implementação representam diferentes grupos encontrados em trabalhos relacionados e destacados na seção 3.4.1 deste documento, tanto com base em redes neurais quanto em árvores.

4.3.5.1 Decision Tree Regressor

O modelo Decision Tree Regressor foi implementado através da aplicação padrão descrita na biblioteca *sklearn*, com o critério de qualidade estabelecido como erro quadrático médio. Foram utilizadas as funções *fit*, para a construção da árvore de decisão a partir do conjunto de treinamento, e *predict* para prever o valor de regressão.

4.3.5.2 Random Forest

O modelo Random Forest também foi implementado seguindo as recomendações da biblioteca *sklearn*. Foram utilizadas 100 árvores na floresta, através do parâmetro *n_estimators*. O

critério de qualidade é o erro quadrático médio.

4.3.5.3 Artificial Neural Network

O modelo ANN foi implementado com base em 70% dos dados para treinamento, organizados de forma aleatória. O algoritmo conta com *early_stopping* (parada precoce), uma abordagem usada para interromper o treinamento quando a métrica parar de melhorar. A aplicação conta com um delta mínimo de 0.001, ou seja, qualquer mudança inferior ao delta não será considerada como melhoria no modelo. Além disso, está definida a paciência do modelo como 20, ou seja, são processadas 20 épocas antes de permitir a parada precoce.

O modelo contou com uma camada densa de 64 neurônios, cuja função de ativação é a ReLU - Ativação Linear Retificada. O otimizador do modelo é o Adam, método de descida gradiente estocástico, baseado na estimativa adaptativa de momentos de primeira e segunda ordem.

Para evitar problemas com *overfitting*, quando o modelo consegue se ajustar com precisão aos dados observados, mas é ineficaz na previsão de novos resultados, foi aplicada a técnica de *dropout*, na qual uma fração de 30% de unidades de entrada de uma camada é eliminada aleatoriamente a cada etapa do treinamento. Dessa forma, a rede possui maiores dificuldades para a formação de padrões de aprendizado frágeis e procura padrões amplos e robustos. Essa abordagem permite com que a previsão não seja feita por uma grande rede e sim por um comitê de redes menores, abordagem semelhante aos modelos de árvore de decisão.

O modelo também contou com a aplicação da camada de normalização de lote, ou *batch normalization*, que ajuda a corrigir situações em que o treinamento da rede está lento ou instável. O *batch size*, ou tamanho do lote, é de 128 e esse parâmetro define o número de amostras que serão propagadas pela rede. A aplicação conta com 100 épocas.

4.3.5.4 XGBoost

A aplicação do modelo XGBoost contou com 1.000 árvores de aumento de gradiente, ou rodadas de reforço, através da definição de *n_estimators*. A taxa de aprendizado utilizada é 0.05 e o objetivo é permitir que as previsões de cada modelo sejam multiplicadas pela taxa antes de serem somadas, assim, cada árvore adicionada ao conjunto interfere menos no resultado, permitindo um número alto de *n_estimators* sem *overfitting*.

Para o XGBoost, a definição de rodadas para parada precoce é de 5, necessitando completar 5 ciclos para permitir interrupções no modelo.

4.4 Discussão

A aplicação dos modelos computacionais seguiu recomendações de boas práticas da biblioteca *sklearn*, além de uma série de testes com diferentes parametrizações para a busca da melhor

combinação de fatores. No caso da rede ANN, por exemplo, foram aplicadas diversas combinações de tamanho de lote, número de camadas e de neurônios, sendo o formato descrito no item 4.3.5.3 o melhor resultado obtido. Os testes podem ser observados no Apêndice B.

5 RESULTADOS

Essa seção visa apresentar os resultados obtidos com a implementação de 4 modelos computacionais para a estimativa de evapotranspiração de referência através de um conjunto limitado de dados climáticos.

5.1 Metodologia dos experimentos

Os experimentos foram realizados através da plataforma *Kaggle* em três páginas diferentes, uma para cada equação de referência. O modelo utilizado como baseline para a implementação é o *Dummy Regressor*, que é uma implementação de regressão que fornece a previsão com base em estratégias simples, sem prestar atenção nos dados de entrada. Esse modelo é frequentemente utilizado como linha de base para a comparação com outros métodos. As métricas escolhidas para a análise da qualidade dos modelos são MAE - Mean Absolute Error e RMSE - Root Mean Square Error, amplamente utilizadas nos estudos relacionados, conforme item 3.4.6 deste trabalho.

5.2 Resultados numéricos

Os modelos de aprendizado de máquina foram implementados e executados com base em três diferentes métodos de comparação: Penman-Monteith, Turc e Makkink, e seus resultados podem ser analisados na Figura 10, conforme parâmetros descritos na seção 4.

Figura 10 – Resultados obtidos

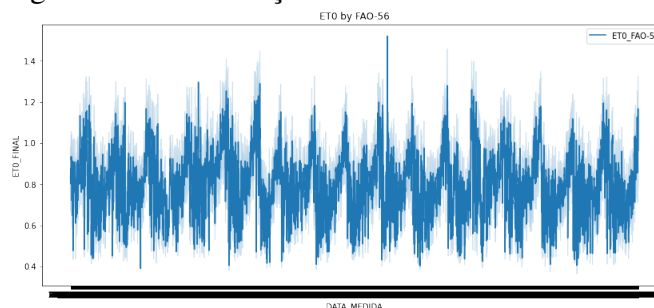
Penman-Monteith			Turc			Makkink		
Modelo	MAE	RMSE	Modelo	MAE	RMSE	Modelo	MAE	RMSE
Baseline - Dummy Regressor	0,1500	0,4220	Baseline - Dummy Regressor	0,1500	0,4202	Baseline - Dummy Regressor	0,4200	0,7056
DecisionTreeRegressor	0,0070	0,0976	DecisionTreeRegressor	0,0027	0,0697	DecisionTreeRegressor	0,0024	0,0618
Random Forest	0,0031	0,0729	Random Forest	0,0012	0,0454	Random Forest	0,0008	0,0431
ANN	0,0112	0,0090	ANN	0,0068	0,0066	ANN	0,0094	0,0124
XGBoost	0,0023	0,0031	XGBoost	0,0011	0,0419	XGBoost	0,0014	0,0019

Fonte: Elaborado pela autora.

É possível perceber que o MAE foi extremamente baixo, apresentando o excelente desempenho dos modelos de aprendizado de máquina na análise de evapotranspiração. A baseline, através do modelo *Dummy Regressor*, apresentou o maior erro, conforme esperado. Os modelos de maior assertividade estão relacionados à aplicação de árvores de decisão, mesmo que com uma margem pequena de diferença das outras implementações.

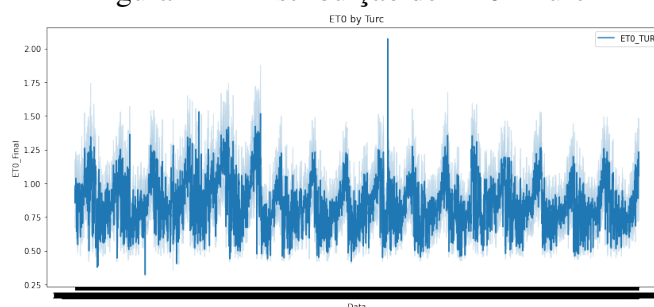
A distribuição dos valores de evapotranspiração de referência para cada uma das equações pode ser analisada de acordo com as Figuras 11, 12 e 13, que contemplam os resultados para todo o período do dataset, entre abril de 2002 e outubro de 2020. É possível observar que o modelo Penman-Monteith permite uma maior suavidade nas diferenças entre os dados, isso ocorre por levar em consideração diversos aspectos no seu cálculo, conforme Apêndice A, não sofrendo tanto o impacto de um único parâmetro.

Figura 11 – Distribuição de ET0 - Penman-Monteith



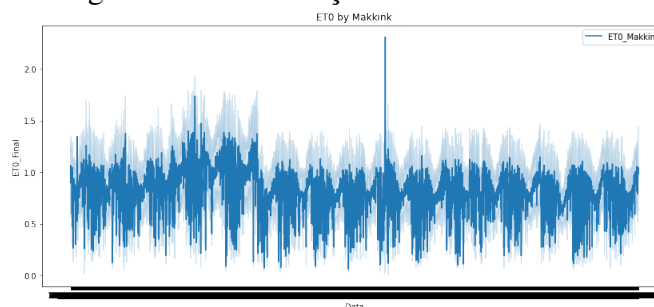
Fonte: Elaborado pela autora.

Figura 12 – Distribuição de ET0 - Turc



Fonte: Elaborado pela autora.

Figura 13 – Distribuição de ET0 - Makkink

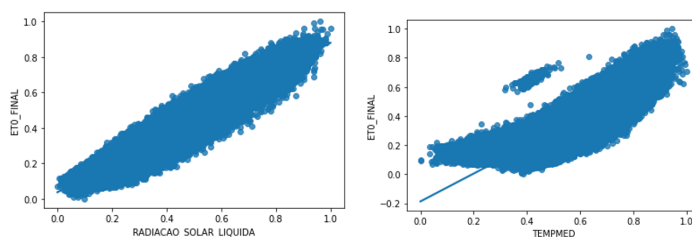


Fonte: Elaborado pela autora.

Além disso, é possível analisar a relação direta das principais features para cada equação, principalmente em relação à temperatura e radiação, aspectos fundamentais para a análise ho-

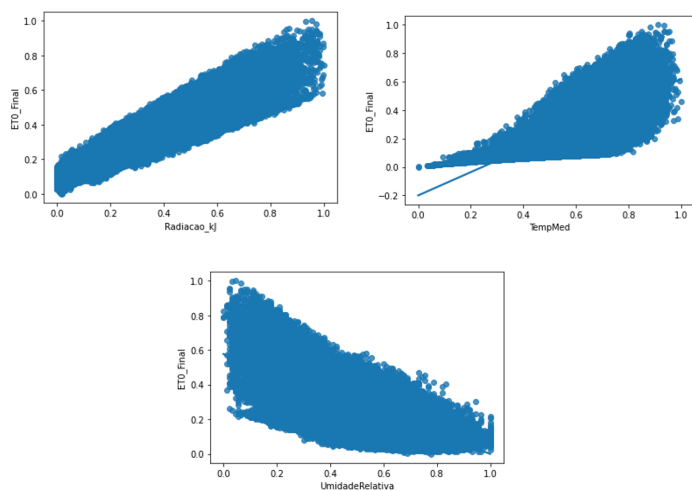
rária de evapotranspiração, conforme Figuras 14, 15 e 16. Essa interpretação permite o entendimento de quais dados meteorológicos são fundamentais para a estimativa de evapotranspiração, reforçando a necessidade de procedimentos como a eliminação de dados cujo sensor de radiação solar apresenta defeito, pois o parâmetro é de extrema importância na análise.

Figura 14 – Impacto dos dados em Penman-Monteith



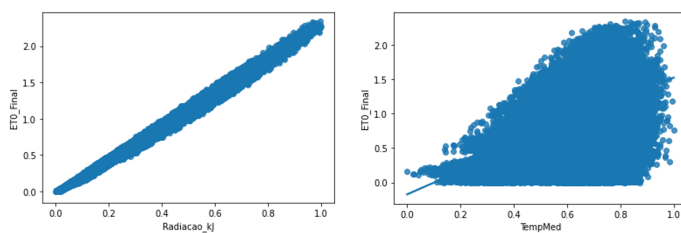
Fonte: Elaborado pela autora.

Figura 15 – Impacto dos dados em Turc



Fonte: Elaborado pela autora.

Figura 16 – Impacto dos dados em Makkink



Fonte: Elaborado pela autora.

5.3 Discussão

Através da aplicação dos modelos de aprendizado de máquina foi possível estimar, de forma assertiva quando comparado com métodos tradicionais e consolidados, a taxa de evapotranspiração. Além disso, foi possível obter o entendimento do efeito dos principais parâmetros climáticos para essa tomada de decisão, sendo a radiação solar totalmente proporcional ao índice de evapotranspiração.

Os modelos desenvolvidos utilizam informações públicas para o treinamento e validação das predições, ou seja, dados disponíveis de forma gratuita para toda a agricultura brasileira. Nesse sentido, os resultados possibilitam a democratização do acesso ao dado de evapotranspiração para todo e qualquer produtor rural. Os produtores rurais poderiam acessar os resultados de evapotranspiração em aplicações simples na internet, através de sua localização para utilização dos dados climáticos adequados. A relevância desse resultado não está apenas na disponibilidade da informação, mas também no impacto que ela pode gerar no aspecto ambiental e econômico. Conforme apresentado, a agricultura utiliza 70% da água mundial e, por falta de informações acessíveis em campo, acaba perdendo cerca de 50% desse total (FAO, 2015). Com a correta interpretação da quantidade de água que está transpirando da planta e evaporando do solo, a decisão de quando e quanto irrigar se torna mais embasada e o produtor rural consegue concentrar o seu esforço na tarefa fundamental da produção de alimentos, enquanto a tecnologia auxilia na tomada de decisão relacionada ao processo de irrigação.

6 CONCLUSÃO E TRABALHOS FUTUROS

Através da implementação do presente trabalho foi possível concluir que métodos computacionais de aprendizado de máquina podem ser utilizados para a análise de evapotranspiração, mesmo com baixa disponibilidade de dados públicos disponíveis. Sendo assim, o objetivo da pesquisa foi alcançado. A aplicação de menor erro ocorre na união da equação de Makkink com o modelo Random Forest. Através da Figura 16 é possível verificar a relação direta entre a radiação e a evapotranspiração, fator que contribui para a facilidade de execução do modelo com base em árvores de decisão.

O impacto do trabalho pode ser separado em três grandes esferas: social, econômica e ambiental. No aspecto social, está sendo democratizado o acesso à informação de evapotranspiração através do uso de dados meteorológicos públicos e de modelos que dispensam uma complexidade matemática para o produtor rural, se atentando ao perfil de escolaridade do público. A esfera econômica está relacionada à redução de gastos com energia elétrica para a irrigação, assim como as possibilidades de aumento de produtividade e segurança de safra pela correta aplicação de água em campo, gerada pelo entendimento dos níveis de evapotranspiração. Por fim, o impacto ambiental está relacionado com a redução do desperdício de água, que hoje ocorre devido ao baixo acesso à tecnologias em campo para entendimento dos momentos de

irrigação.

O trabalho foi aplicado apenas com dataset do oeste baiano, região de papel fundamental na agricultura irrigada brasileira. Como trabalhos futuros, recomenda-se a expansão da pesquisa para a modelagem em diferentes regiões, com climas de maior variabilidade e incidência de chuva. Além disso, há a oportunidade de implementação e teste de novos modelos de aprendizado de máquina, gerando uma comparação aos resultados encontrados. Ainda como trabalhos futuros, é possível realizar a validação das taxas de evapotranspiração durante uma safra agrícola, com acompanhamento de produtores rurais, para o levantamento dos impactos econômicos dos modelos na tomada de decisão.

É importante ressaltar o papel da tecnologia no agronegócio, setor fundamental para o desenvolvimento sustentável e econômico do Brasil e do mundo. Soluções como as apresentadas no presente trabalho aproximam a realidade da computação da realidade do campo e permitem o desenvolvimento de todos os setores.

Referências

AGATONOVIC-KUSTRIN, S.; BERESFORD, R. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. **Journal of Pharmaceutical and Biomedical Analysis**, v. 22, n. 5, p. 717–727, 2000. ISSN 07317085.

ANA. Atlas irrigação: uso da água na agricultura irrigada. 2017.

ANTONOPOULOS, V. Z.; ANTONOPOULOS, A. V. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. **Computers and Electronics in Agriculture**, Elsevier B.V., v. 132, p. 86–96, 2017. ISSN 01681699. Disponível em: <<http://dx.doi.org/10.1016/j.compag.2016.11.011>>.

BROWNLEE, J. A gentle introduction to xgboost for applied machine learning. Machine Learning Mastery, 2016.

CEPEA. PIB DO AGRONEGÓCIO BRASILEIRO. 2020.

DAMACENO, S. S.; VASCONCELOS, R. O. Inteligência artificial: uma breve abordagem sobre seu conceito real e o conhecimento popular. p. 11–16, 2018. Disponível em: <<https://periodicos.set.edu.br/cadernoexatas/article/view/5729>>.

DANIEL, D. F. et al. Evapotranspiração e produtividade de milho safrinha consorciado com crotalária. **Research, Society and Development**, v. 9, n. 8, 2020. ISSN 1098-6596.

ELAVARASAN, D. et al. Forecasting yield by integrating agrarian factors and machine learning models: A survey. **Computers and Electronics in Agriculture**, Elsevier, v. 155, n. August, p. 257–282, 2018. ISSN 01681699. Disponível em: <<https://doi.org/10.1016/j.compag.2018.10.024>>.

FAO. Se o atual ritmo de consumo continuar, em 2050 mundo precisará de 60% mais alimentos e 40% mais água. 2015.

FERREIRA, L. B.; CUNHA, F. F. da. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. **Computers and Electronics in Agriculture**, Elsevier, v. 178, n. May, p. 105728, 2020. ISSN 01681699. Disponível em: <<https://doi.org/10.1016/j.compag.2020.105728>>.

GHOLAMY, A.; KREINOVICH, V.; KOSHELEVA, O. Why 70/30 or 80/20 Relation Between Training and Testing Sets : A Pedagogical Explanation. **Departmental Technical Reports (CS)**, p. 1–6, 2018.

GUPTA, M. ML | feature scaling – part 2. GeeksforGeeks, 2021.

IBGE, I. B. de Geografia e E. **Censo Agropecuário 2017 - Resultados Definitivos**. 2017.

KAYA, Y. Z.; MAMAK, M.; UNES, F. Evapotranspiration Prediction Using M5T Data Mining Method. **International Journal of Advanced Engineering Research and Science**, v. 3, n. 12, p. 225–229, 2016. ISSN 23496495.

LUCAS, P. de Oliveira e. Previsão de Séries Temporais de Evapotranspiração de Referência com Redes Neurais Convolucionais. 2019.

LUCAS, R. P. O. et al. USING ARTIFICIAL NEURAL NETWORK TO ESTIMATE REFERENCE EVAPOTRANSPIRATION. p. 229–240, 2018.

MINACAPILLI, M. et al. Using scintillometry to assess reference evapotranspiration methods and their impact on the water balance of olive groves. **Agricultural Water Management**, Elsevier B.V., v. 170, p. 49–60, 2016. ISSN 18732283. Disponível em: <<http://dx.doi.org/10.1016/j.agwat.2015.12.004>>.

MORAES, S. O.; RENÊ, C.; JÚNIOR, A. B. **Evapotranspiração e sua medida por lisímetro de pesagem**. [S.l.], 2009.

OLIVEIRA, P. L. de et al. **AVALIAÇÃO METROLÓGICA DE UM SISTEMA DE CALIBRAÇÃO INDOOR PARA PIRANÔMETROS BASEADO EM UM SIMULADOR SOLAR**. 2013. 17-24 p.

ONU, O. das N. U. Transformando Nosso Mundo : A Agenda 2030 para o Desenvolvimento Sustentável. p. 1–41, 2015.

PEREIRA VILLA NOVA, S. **evapo(transpi)ração**. 1997.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. In: **Information and Software Technology**. [S.l.]: Elsevier, 2015. v. 64, p. 1–18. ISSN 09505849.

RIBEIRO, J.; MARINHO, D.; ESPINOSA, J. **Agricultura 4.0: Desafios à produção de alimentos e inovações tecnológicas**. 2017.

RODRIGUES, C. M. Capítulo 8 Evaporação e evapotranspiração. p. 215–234, 2017.

SANTIAGO, A. V. **EVAPOTRANSPIRAÇÃO DE REFERÊNCIA MEDIDA POR LISÍMETRO DE PESAGEM E ESTIMADA POR PENMAN-MONTEITH (FAO-56), NAS ESCALAS MENSAL E DECENDIAL**. [S.l.], 2001.

SILVA, G. **DETECÇÃO E CONTAGEM DE PLANTAS UTILIZANDO TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING**. Tese (DEPARTAMENTO DE ENGENHARIA ELÉTRICA E ELETRÔNICA) — UNIVERSIDADE FEDERAL DE SANTA CATARINA, 2017. Disponível em: <lib.unnes.ac.id/6871/1/8479.pdf{\%}0Ahttp://www.albayan>.

TORRES, A. B. et al. Multilevel data fusion for the internet of things in smart agriculture. **Computers and Electronics in Agriculture**, Elsevier, v. 171, n. November 2019, p. 105309, 2020. ISSN 01681699. Disponível em: <<https://doi.org/10.1016/j.compag.2020.105309>>.

WHO, W. H. O. Drinking-Water. 2019.

YUSUF, H. A.; ABIDAKUN, O. J.; MAKINDE, S. B. Evapotranspiration Estimation Using Artificial Neural Network over South-Western Nigeria. **Journal of Engineering Research and Reports**, v. 17, n. 1, p. 16–35, 2020.

ZHU, B. et al. Hybrid particle swarm optimization with extreme learning machine for daily reference evapotranspiration prediction from limited climatic data. **Computers and Electronics in Agriculture**, Elsevier, v. 173, n. April, p. 105430, 2020. ISSN 01681699. Disponível em: <<https://doi.org/10.1016/j.compag.2020.105430>>.

APÊNDICE A – CÁLCULO DE EVAPOTRANSPIRAÇÃO ATRAVÉS DE FAO-56

A realização do cálculo de evapotranspiração através do método Penman-Monteith, ou FAO-56, possui um nível de complexidade matemática avançada quando trabalhado através de dados obtidos por estações meteorológicas convencionais. Para essa finalidade, são necessários 19 passos de cálculo. Os cálculos são:

1. Dia juliano

$$J = DM - PDA \quad (8)$$

Onde:

J = Dia Juliano (dias)

DM = Dia da medida (dia/mês/ano)

PDA = Primeira data do ano (dia/mês/ano)

2. Declinação do Sol

$$\delta = 0,4093 \times \text{sen}\left(\frac{2\pi}{365}J - 1,405\right) \quad (9)$$

Onde:

δ = Declinação do sol (radianos)

J = Dia Juliano (dias)

π = Constante pi = 3,1415

3. Inverso da distância relativa entre o sol e a terra

$$d_r = 1 + 0,033 \times \cos\left(\frac{2\pi}{365}J\right) \quad (10)$$

Onde:

d_r = Inverso da distância relativa entre o sol e a terra

J = Dia Juliano (dias)

π = Constante pi = 3,1415

4. Transformação de latitude em radianos

$$\varphi = (\text{sinal}) \times \text{rad}\left(\text{graus} + \frac{\text{min}}{60} + \frac{\text{seg}}{3600}\right) \quad (11)$$

Onde:

φ = Latitude do ponto medido (radianos)
 $sin\alpha = -1$ para hemisfério sul e $+1$ para hemisfério norte
 $graus$ = Latitude, valor apenas dos graus
 min = Latitude, valor apenas dos minutos
 seg = Latitude, valor apenas dos segundos

5. Ângulo solar do por do sol

$$\omega_s = \arccos[-\tan\varphi \times \tan\delta] \quad (12)$$

Onde:

ω_s = Ângulo solar do por do sol (radianos)
 φ = Latitude (radianos)
 δ = Declinação do sol (radianos)

6. Radiação solar extraterreste

$$R_a = 37,6 \times d_r(\omega_s \times \sin\varphi \times \sin\delta + \cos\varphi \times \cos\delta \times \sin\omega_s) \quad (13)$$

Onde:

R_a = Radiação solar extraterreste (MJ/m²dia)
 r = Inverso da distância relativa entre o sol e a terra
 ω_s = Ângulo solar do por do sol (radianos)
 φ = Latitude (radianos)
 δ = Declinação do sol (radianos)

7. Transformação da radiação solar global em MJ/m²dia

$$R_s = 86,4 \times \frac{KW}{m^2} \quad (14)$$

Onde:

R_s = Radiação solar global (MJ/m²dia)
 $\frac{KW}{m^2}$ = Valor obtido pelo parâmetro da estação

8. Radiação solar Global em dia de sol claro

$$R_{so} = [0,75 + 0,00002 \times (h + z)] \times R_a \quad (15)$$

Onde:

- R_{so} = Radiação solar global em dia de sol claro (MJ/m²dia)
 R_a = Radiação solar extraterrestre (MJ/m²dia)
 h = Altura da medida do vento (m)
 z = Altitude (m)

9. Pressão de saturação de vapor d'água

$$e_s = \frac{0,6108 \times \exp\left[\frac{17,27 \times T_{min}}{T_{min} + 237,3}\right] + 0,6108 \times \exp\left[\frac{17,27 \times T_{max}}{T_{max} + 237,3}\right]}{2} \quad (16)$$

Onde:

- e_s = Pressão de saturação de vapor d'água (KPa)
 T_{min} = Temperatura mínima (°C)
 T_{max} = Temperatura máxima (°C)

10. Pressão parcial de vapor d'água

$$e_a = \frac{0,6108 \times \exp\left[\frac{17,27 \times T_{min}}{T_{min} + 237,3}\right] \times \left(\frac{UR_{max}}{100}\right) + 0,6108 \times \exp\left[\frac{17,27 \times T_{max}}{T_{max} + 237,3}\right] \times \left(\frac{UR_{min}}{100}\right)}{2} \quad (17)$$

Onde:

- e_a = Pressão parcial de vapor d'água (KPa)
 T_{min} = Temperatura mínima (°C)
 T_{max} = Temperatura máxima (°C)
 UR_{min} Umidade relativa mínima (%)
 UR_{max} Umidade relativa máxima (%)

11. Radiação líquida de ondas longas

$$R_{nl} = \sigma \times \left(\frac{(T_{max} + 273,16)^4 + (T_{min} + 273,16)^4}{2} \right) \times (0,34 - 0,14\sqrt{e_a}) \times \left(1,35 \times \frac{R_s}{R_{so}} - 0,35 \right) \quad (18)$$

Onde:

- R_{nl} = Radiação líquida de ondas longas (MJ/m²dia)
 σ = Constante de Stefan-Boltzmann = 4,903.10⁻⁹ MJ/m²dia
 T_{min} = Temperatura mínima (°C)
 T_{max} = Temperatura máxima (°C)
 e_a = Pressão parcial de vapor d'água (KPa)

R_s = Radiação solar global (MJ/m²dia)

R_{so} = Radiação solar global em dias de sol claro (MJ/m²dia)

12. Radiação líquida de ondas curtas

$$R_{ns} = (1 - \alpha) \times R_s \quad (19)$$

Onde:

R_{ns} = Radiação líquida de ondas curtas (MJ/m²dia)

R_s = Radiação solar global (MJ/m²dia)

α = Albedo, coeficiente de reflexão da cultura (adimensional).

Para cálculo de Et0 utiliza-se albedo da cultura da grama, que é 0,23.

13. Radiação líquida

$$R_n = R_{ns} - R_{nl} \quad (20)$$

Onde:

R_n = Radiação líquida (MJ/m²dia)

R_{ns} = Radiação líquida de ondas curtas (MJ/m²dia)

R_{nl} = Radiação líquida de ondas longas (MJ/m²dia)

14. Temperatura média

$$T_{med} = \frac{T_{max} + T_{min}}{2} \quad (21)$$

Onde:

T_{med} = Temperatura média (°C)

T_{max} = Temperatura máxima (°C)

T_{min} = Temperatura mínima (°C)

15. Inclinação da curva de pressão de vapor d'água

$$\Delta = \frac{4098 \times (0,6108 \times \exp(\frac{17,27 \times T_{med}}{T_{med} + 237,3}))}{(T_{med} + 237,3)^2} \quad (22)$$

Onde:

Δ = Declividade da curva de pressão de vapor (KPa °C⁻¹)
 T_{med} = Temperatura média (°C)

16. Pressão atmosférica

$$P_a = 101,3 \times \left(\frac{293 - 0,0065 \times z}{293} \right)^{5,26} \quad (23)$$

Onde:

P_a = Pressão atmosférica (KPa)
 z = Altitude (m)

17. Constante psicométrica

$$\gamma = 0,000665 \times P_a \quad (24)$$

Onde:

γ = Constante psicométrica (KPa °C⁻¹)
 P_a = Pressão atmosférica (KPa)

18. Velocidade do vento a 2 m/s

$$u_2 = u_z \left(\frac{4,87}{\ln(67,8 \times z - 5,42)} \right) \quad (25)$$

Onde:

u_2 = Velocidade média do vento a 2 metros de altura (m s⁻¹)
 u_z = Velocidade média do vento obtida na estação meteorológica (m s⁻¹)
 z = Altura da medição da velocidade do vento (m)

19. Evapotranspiração de referência através de FAO-56

$$ET_0 = \frac{0,408 \times \Delta(R_n - G) + \gamma \frac{37}{T_{med} + 237} \times u_2(e_s - e_a)}{\Delta + \gamma(1 + 0,34 \times u_2)} \quad (26)$$

Onde:

ET_0 = Evapotranspiração de referência (mm h⁻¹)
 Δ = Inclinação da curva de pressão de vapor (KPa °C⁻¹)
 R_n = Saldo de radiação (MJ m²h⁻¹)
 G = Fluxo de calor no solo (MJ m²h⁻¹)

Nesse caso, considera-se 0.

γ = Constante psicométrica (KPa °C⁻¹)

T_{med} = Temperatura média (°C)

u_2 = Velocidade do vento a 2 metros de altura (m 2⁻¹)

e_s = Pressão de saturação do vapor d'água (KPa)

e_a = Pressão de vapor d'água (KPa)

APÊNDICE B – ROTINA DE TESTES DO MODELO ANN

Para a validação dos parâmetros do modelo ANN, foram realizadas uma série de testes dentro do código de referência com cálculo Turc. A Figura 17 apresenta os resultados obtidos e a escolha pela modelagem atual.

Figura 17 – Variação de parâmetros para ANN

Batch Size	Otimizador	Neurônios Layer 1	Neurônios Layer 2	Neurônios Layer 3	MAE
32	ADAM	32	16	-	0.0107
32	ADAM	32	-	-	0.0106
32	ADAM	32	32	-	0.0168
32	ADAM	32	32	32	0.0145
32	ADAM	64	32	32	0.0088
32	ADAM	64	32	16	0.0116
32	ADAM	64	16	16	0.0103
32	ADAM	128	64	32	0.0105
32	ADAM	128	32	32	0.0091
32	ADAM	128	64	64	0.01
32	ADAM	128	128	128	0.0091
32	NADAM	32	-	-	0.0105
32	RMSPROP	32	-	-	0.0071
64	ADAM	32	16	-	0.0224
64	ADAM	32	32	-	0.0232
64	ADAM	32	-	-	0.0134
64	RMSPROP	32	-	-	0.0099
64	NADAM	32	-	-	0.0202
64	NADAM	32	32	-	0.0171
64	NADAM	64	32	-	0.0069
128	ADAM	32	16	-	0.0387
128	ADAM	32	-	-	0.01
128	ADAM	32	32	-	0.0205
128	ADAM	32	32	32	0.0085
128	ADAM	64	32	32	0.0069
128	ADAM	64	-	-	0.0068
128	NADAM	64	-	-	0.1269
256	ADAM	32	-	-	0.0185
256	ADAM	32	16	-	0.0089
256	ADAM	32	32	-	0.031
256	ADAM	16	-	-	0.0186
256	NADAM	16	-	-	0.0205
512	ADAM	16	-	-	0.0345
512	NADAM	64	-	-	0.0075
512	ADAM	64	-	-	0.0074
512	ADAM	64	32	-	0.0157
512	RMSPROP	64	-	-	0.009
512	NADAM	64	64	-	0.018
512	NADAM	64	32	16	0.0102
1024	ADAM	64	-	-	0.007
1024	ADAM	32	-	-	0.0084
1024	ADAM	64	32	-	0.0306

Fonte: Elaborado pela autora.