



UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS
CURSO DE ENGENHARIA CIVIL

JOSÉ KERLLY SOARES DE ARAÚJO

AVALIAÇÃO DE MODELOS DE PREVISÃO DE VAZÃO BASEADOS EM
APRENDIZAGEM DE MÁQUINAS - RESERVATÓRIO FRONTEIRAS, CRATEÚS -
CE

CRATEÚS
2023

JOSÉ KERLLY SOARES DE ARAÚJO

AVALIAÇÃO DE MODELOS DE PREVISÃO DE VAZÃO BASEADOS EM
APRENDIZAGEM DE MÁQUINAS - RESERVATÓRIO FRONTEIRAS, CRATEÚS - CE

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Civil da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Engenharia Civil.

Orientador: Prof. Dr. Alan Michell Barros Alexandre.

CRATEÚS

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

A689a Araújo, José Kerlly Soares de.
Avaliação de Modelos de Previsão de Vazão baseados em Aprendizagem de Máquinas - Reservatório Fronteiras, Crateús-CE / José Kerlly Soares de Araújo. – 2023.
89 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Crateús, Curso de Engenharia Civil, Crateús, 2023.
Orientação: Prof. Dr. Alan Michell Barros Alexandre.

1. Previsão de vazão. 2. Índices TSM. 3. Aprendizado de máquina. 4. Redes neurais artificiais. 5. Regressão linear. I. Título.

CDD 620

JOSÉ KERLLY SOARES DE ARAÚJO

AVALIAÇÃO DE MODELOS DE PREVISÃO DE VAZÃO BASEADOS EM
APRENDIZAGEM DE MÁQUINAS - RESERVATÓRIO FRONTEIRAS, CRATEÚS - CE

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Civil da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Engenharia Civil.

APROVADA EM: __/__/____.

BANCA EXAMINADORA

Prof. Dr. Alan Michell Barros Alexandre (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dra. Carla Beatriz Costa de Araújo
Universidade Federal do Ceará (UFC)

Prof. Ma. Tatiane Lima Batista
Universidade Federal do Ceará (UFC)

A minha família, razão de minha existência.

A Deus.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, fonte de toda sabedoria e força, por me guiar em todos os momentos e me dar a perseverança necessária para concluir esta jornada.

Aos meus queridos pais, Cleyde e José, meu alicerce e inspiração, agradeço pelo amor incondicional, apoio e incentivo constante. Vocês foram essenciais em cada passo desta caminhada.

À minha namorada, Ana Isabel, agradeço pelo apoio emocional, compreensão e paciência nos momentos mais desafiadores. Sua presença foi um farol de motivação e conforto.

Aos meus amigos da graduação, Francisco Jonatas, Euller Loiola, Mateus Ferreira e Antonio Serafim, minha profunda gratidão por estarem sempre ao meu lado durante essa jornada acadêmica. Vocês ofereceram palavras de incentivo, momentos de descontração e companheirismo que foram essenciais para manter o equilíbrio e a motivação durante este processo. Além disso, gostaria de estender meus agradecimentos aos amigos que fiz fora do ambiente acadêmico. Cada um de vocês, com sua maneira única, contribuiu para meu crescimento pessoal e profissional nessa etapa importante da minha vida.

Um agradecimento especial ao meu orientador, Alan Michell, cuja sabedoria, paciência e orientação foram cruciais para o desenvolvimento e aprimoramento deste trabalho. Seu apoio e conhecimento foram fundamentais para que eu alcançasse este objetivo.

Por fim, à banca examinadora, agradeço pelas valiosas considerações e feedbacks que enriqueceram significativamente este trabalho.

Este projeto é um marco importante na minha vida acadêmica e pessoal, e sou grato a todos que fizeram parte desta jornada.

“Feliz aquele que transfere o que sabe e aprende
o que ensina.”

Cora Coralina

RESUMO

O semiárido do Nordeste brasileiro é uma área caracterizada por condições climáticas e hidrológicas adversas, como baixa pluviosidade e irregularidade das chuvas. Esses fatores, juntamente com a geologia local, dificultam perenização natural dos rios e a recarga de aquíferos. Como resposta, a construção de reservatórios artificiais, exemplificada pelo açude Fronteiras no Ceará, é crucial para a sustentabilidade hídrica da região. Dessa forma, a capacidade de previsão de vazões afluentes aos reservatórios é um elemento central para a gestão eficiente dos recursos hídricos, na qual as previsões e incertezas devem ser analisadas no funcionamento e planejamento do sistema. Em função disso, o presente trabalho tem como objetivo principal avaliação da previsibilidade do regime vazão anual afluente ao Reservatório Fronteiras, a partir da variabilidade dos índices de Temperatura de Superfície do Mar (TSM) utilizando técnicas de redes neurais artificiais e modelos de regressão linear. Para isso foram utilizados como períodos preditores os trimestres Setembro-Outubro-Novembro (SON), Junho-Julho-Agosto (JJA), Março-Abril-Maio (MAM) e Dezembro-Janeiro-Fevereiro (DJF), em que os diferentes índices TSM são ranqueados por ordem de relevância através de técnicas de aprendizado de máquinas (*Random Forest*, *Select K Best* e *XGBoost*). Em seguida foram definidos modelos de previsão utilizando os métodos *Multilayer Perceptron* (MLP), *Long Short-Term Memory* (LSTM), Regressões Lasso, Ridge e Regressão Linear Múltipla (RLM). Na análise dos resultados, os fatores mais significativos identificados foram o Dipolo do Atlântico Tropical (TAD) e as anomalias de TSM associadas a ele, especificamente o Atlântico Tropical Norte e o Atlântico Tropical Sul. Os modelos mais eficazes para prever a vazão pertencem ao período preditor SON, destacando-se como melhor modelo o MLP-M7, que obteve um NSE (Nash-Sutcliffe Efficiency) de 0,329 e uma correlação de 0,638 durante a fase de validação. Por fim, apesar dos modelos apresentarem correlações satisfatórias, os valores de NSE foram considerados moderadamente baixos.

Palavras-chave: Previsão de vazão. Índices TSM. Aprendizado de máquina. Redes neurais artificiais. Regressão linear.

ABSTRACT

The semi-arid region of Northeastern Brazil is characterized by challenging climatic and hydrological conditions, such as low rainfall and irregular rain patterns. These factors, combined with the local geology, hinder the natural perennial flow of rivers and the recharge of aquifers. In response, the construction of artificial reservoirs, exemplified by the Fronteiras reservoir in Ceará, is crucial for the region's water sustainability. Consequently, predicting inflow to these reservoirs is a key element in the efficient management of water resources, where forecasts and uncertainties must be analyzed for system operation and planning. This study primarily aims to evaluate the predictability of the annual inflow regime to the Fronteiras Reservoir, based on the variability of Sea Surface Temperature (SST) indices using artificial neural networks and linear regression models. Predictor periods such as September-October-November (SON), June-July-August (JJA), March-April-May (MAM), and December-January-February (DJF) were used, ranking different SST indices by relevance using machine learning techniques (Random Forest, Select K Best, and XGBoost). Prediction models were then defined using Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Lasso and Ridge Regressions, and Multiple Linear Regression (MLR). In the results, the most significant factors identified were the Tropical Atlantic Dipole (TAD) and SST anomalies associated with it, specifically the North and South Tropical Atlantic. The most effective models for predicting flow belong to the SON predictor period, with the MLP-M7 model standing out, achieving a Nash-Sutcliffe Efficiency (NSE) of 0.329 and a correlation of 0.638 during the validation phase. Finally, despite satisfactory correlations, the NSE values were considered moderately low.

Keywords: Flow prediction. SST indices. Machine learning. Artificial neural networks. Linear regression.

LISTA DE SIGLAS

AM	Aprendizado de máquina
AMO	Oscilação Multidecadal do Atlântico
ANA	Agência Nacional de Águas e Saneamento Básico
COGERH	Companhia de Gestão de Recursos Hídricos
DJF	Período de dezembro a fevereiro
ENOS	El Niño – Oscilação Sul
ESRL	Earth System Research Laboratory
IBGE	Instituto Brasileiro de Geografia e Estatística
IPECE	Instituto de Pesquisa e Estratégia Econômica do Ceará
LSTM	Long Short Term Memory
MAM	Período de março a maio
NAO	Oscilação do Atlântico Norte
NEB	Nordeste brasileiro
NINO1.2	Niño 1+2
NINO3	Niño 3
NINO3.4	Niño 3.4
NINO4	Niño 4
NOAA	National Oceanic and Atmospheric Administration
MLP	Multilayer Perceptron
PDO	Oscilação Decadal do Pacífico
RHSC	Região Hidrográfica dos Sertões de Crateús
RNA	Redes Neurais Artificiais
RNR	Rede Neural Recorrente
RLM	Regressão Linear Múltipla
SRH	Secretaria de Recursos Hídricos
TAD	Dipolo do Atlântico Tropical
TNA	Atlântico Norte Tropical
TSA	Atlântico Sul Tropical
TSM	Temperatura de Superfície do Mar
ZCIT	Zona de Convergência Intertropical

LISTA DE FIGURAS

Figura 1 - Estrutura do modelo matemático de um neurônio artificial	27
Figura 2 - Estrutura de uma Rede Neural MultiLayer Perceptron	28
Figura 3 - Representação de uma célula LSTM.	33
Figura 4 – LSTM replicada ao longo do tempo.	33
Figura 5 – Bacia Hidrográfica do Reservatório Fronteiras.....	38
Figura 6 – Estratégia para definição dos modelos de previsão.....	42
Figura 7 - <i>Boxplot</i> da importância dos índices TSM no período preditor SON (<i>Random Forest</i>)	50
Figura 8 - Importância dos índices TSM no período preditor SON (<i>Select K Best</i>)	51
Figura 9 - Importância dos índices TSM no período preditor SON (<i>XGBoost</i>)	51
Figura 10 - Métricas de Avaliação do Modelo MLP com Preditor SON (NSE e Correlação) ..	53
Figura 11 - Métricas de Avaliação do Modelo LSTM com Preditor SON (NSE e Correlação)	53
Figura 12 - Métricas de Avaliação do Modelo LASSO com Preditor SON (NSE e Correlação)	54
Figura 13 - Métricas de Avaliação do Modelo RIDGE com Preditor SON (NSE e Correlação)	54
Figura 14 - Métricas de Avaliação do Modelo RLM com Preditor SON (NSE e Correlação) ..	54
Figura 15 - <i>Boxplot</i> da importância dos índices TSM no período preditor JJA (<i>Random Forest</i>)	56
Figura 16 - Importância dos índices TSM no período preditor JJA (<i>Select K Best</i>)	57
Figura 17 - Importância dos índices TSM no período preditor JJA (<i>XGBoost</i>)	57
Figura 18 - Métricas de Avaliação do Modelo MLP com Preditor JJA (NSE e Correlação) ..	59
Figura 19 - Métricas de Avaliação do Modelo LSTM com Preditor JJA (NSE e Correlação) ..	59
Figura 20 - Métricas de Avaliação do Modelo LASSO com Preditor JJA (NSE e Correlação)	60
Figura 21 - Métricas de Avaliação do Modelo RIDGE com Preditor JJA (NSE e Correlação)	60
Figura 22 - Métricas de Avaliação do Modelo RLM com Preditor JJA (NSE e Correlação) ..	60
Figura 23 - <i>Boxplot</i> da importância dos índices TSM no período preditor MAM (<i>Random Forest</i>)	62
Figura 24 - Importância dos índices TSM no período preditor MAM (<i>Select K Best</i>)	63

Figura 25 - Importância dos índices TSM no período preditor MAM (<i>XGBoost</i>).....	63
Figura 26 - <i>Boxplot</i> da importância dos índices TSM no período preditor DJF (<i>Random Forest</i>)	67
Figura 27 - Importância dos índices TSM no período preditor DJF (<i>Select K Best</i>).....	68
Figura 28 - Importância dos índices TSM no período preditor DJF (<i>XGBoost</i>)	68
Figura 29 – Comparação entre os Modelos de SON (NSE - validação)	72
Figura 30 – Comparação entre os Modelos de SON (R - Validação)	73
Figura 31 – Comparação entre os modelos de JJA (NSE - validação).....	75
Figura 32 – Comparação entre os modelos de JJA (R - validação).....	75

LISTA DE QUADROS

Quadro 1 - Índices TSM utilizados	39
Quadro 2 – Metodologia para o ranqueamento unificado	41
Quadro 3 – Funcionamento método de Copeland	44
Quadro 4 – Hiperparâmetros da rede neural LSTM.....	47
Quadro 5 – <i>Ranking</i> Copeland para período SON	52
Quadro 6 – <i>Ranking</i> Copeland para período JJA	58
Quadro 7 – <i>Ranking</i> Copeland para período MAM	64
Quadro 8 – <i>Ranking</i> Copeland para período DJF.....	69

LISTA DE TABELAS

Tabela 1 - Avaliação do NSE Para o Preditor MAM	65
Tabela 2 - Avaliação da Correlação de Pearson Para o Preditor MAM	65
Tabela 3 - Avaliação do NSE Para o Preditor DJF	70
Tabela 4 - Avaliação da Correlação de Pearson Para o Preditor DJF	70
Tabela 5 – Variação Percentual de Modelos de SON vs. M1 (NSE - Validação)	73
Tabela 6 – Variação Percentual de Modelos de SON vs. M1 (R - Validação)	73
Tabela 7 – Variação Percentual dos Modelos de JJA vs. M1 (NSE - Validação).....	76
Tabela 8 – Variação Percentual dos Modelos de JJA vs. M1 (R - Validação).....	76

SUMÁRIO

1	INTRODUÇÃO	17
1.1	Objetivos	20
1.1.1	<i>Objetivo geral</i>	20
1.1.2	<i>Objetivos específicos</i>	20
1.2	Roteiro de trabalho	20
2	REVISÃO BIBLIOGRÁFICA	21
2.1	Índices de Temperatura de Superfície do Mar	21
2.1.1	<i>Dipolo do Atlântico</i>	21
2.1.2	<i>ENOS - El Niño e Oscilação Sul</i>	22
2.1.3	<i>Oscilação Decadal do Pacífico</i>	22
2.1.4	<i>Oscilação Multidecadal do Atlântico</i>	23
2.1.5	<i>Oscilação do Atlântico Norte</i>	23
2.2	Aprendizado de máquina (<i>Machine Learning</i>)	23
2.3	Seleção de Características (<i>Feature Selection</i>)	24
2.3.1	<i>Random Forest</i>	25
2.3.2	<i>Seleção baseada em filtro (Filter-based selection)</i>	25
2.3.3	<i>XGBoost</i>	25
2.4	Redes neurais artificiais	26
2.4.1	<i>Introdução a redes neurais</i>	26
2.4.2	<i>Rede Neural MultiLayer perceptron</i>	28
2.4.3	<i>Rede Neural LSTM</i>	31
2.4.4	<i>Ajuste de hiperparâmetros</i>	33
2.5	Regressão Linear Múltipla	34
2.5.1	<i>Regressão Linear Múltipla Tradicional</i>	34
2.5.2	<i>Regressão Lasso e Ridge</i>	36
3	METODOLOGIA	37
3.1	Região de estudo	37
3.2	<i>Levantamento dos dados</i>	39
3.3	Estratégia metodológica	40
3.4	Classificação por ordem de importância dos índices climáticos	42
3.4.1	<i>Random Forest</i>	42
3.4.2	<i>Select K Best</i>	43

3.4.3	<i>XGboost</i>	43
3.4.4	<i>Unificação dos Rankings pelo Método de Copeland</i>	44
3.5	Modelos de Previsão	44
3.5.1	<i>Pré-Processamento dos dados</i>	45
3.5.2	<i>Estruturação dos modelos e determinação de hiperparâmetros</i>	45
3.5.3	<i>Multi Layer Perceptron</i>	45
3.5.4	<i>Long Short Term Memory</i>	46
3.5.5	<i>Regressões Lasso, Ridge e Linear Múltipla</i>	47
3.5.6	<i>Treinamento e validação</i>	48
3.5.7	<i>Métricas de avaliação</i>	48
4	RESULTADOS E DISCUSSÕES	50
4.1	Análise dos resultados para o período preditor SON	50
4.1.1	<i>Análise da importância dos índices TSM para o período preditor SON</i>	50
4.1.2	<i>Ranqueamento unificado pelo método Copeland para o Preditor SON</i>	52
4.1.3	<i>Modelos de previsão de vazão – SON</i>	53
4.2	Análise dos resultados para o período preditor JJA	56
4.2.1	<i>Análise da importância dos índices TSM para o período preditor JJA</i>	56
4.2.2	<i>Ranqueamento unificado pelo método Copeland para o Preditor JJA</i>	58
4.2.3	<i>Modelos de previsão de vazão – JJA</i>	59
4.3	Análise dos resultados para o período preditor MAM	62
4.3.1	<i>Análise da importância dos índices TSM para o período preditor MAM</i>	62
4.3.2	<i>Ranqueamento unificado pelo método Copeland para o Preditor MAM</i>	64
4.3.3	<i>Modelos de previsão de vazão – MAM</i>	65
4.4	Análise dos resultados para o período preditor DJF	67
4.4.1	<i>Análise da importância dos índices TSM para o período preditor DJF</i>	67
4.4.2	<i>Ranqueamento unificado pelo método Copeland para o Preditor DJF</i>	69
4.4.3	<i>Modelos de previsão de vazão – DJF</i>	70
4.5	Comparação e Síntese dos Resultados	72
5	CONCLUSÃO	78
	REFERÊNCIAS	80

1 INTRODUÇÃO

O semiárido do Nordeste Brasileiro (NEB) apresenta particularidades climáticas marcantes e complexas, como a predominância de baixa pluviosidade média anual, irregularidade das chuvas ao longo dos anos, e grande concentração em poucos meses (Monteiro; Zanella; Pinheiro, 2021; Silva; Farias, 2021). A região é caracterizada por elevadas taxas de evaporação, extensos períodos de insolação, além de uma estrutura geológica cristalina que dificulta o reabastecimento dos aquíferos subterrâneos (Novais; Carvalho Júnior; Oliveira, 2022). Dessa forma, predominam rios intermitentes em uma extensão significativa do território semiárido, evidenciando a complexidade do sistema hídrico regional e altos níveis de criticidades (Agência Nacional de Águas, 2012).

O estado do Ceará está localizado na porção setentrional do NEB, predominantemente na parcela semiárida, o que o torna suscetível a períodos de seca prolongados (Costa; Silva, 2017). A composição geológica do estado é majoritariamente caracterizada por rochas cristalinas, que representam aproximadamente 85% de sua área total, conforme destacado pelo Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE) em 2021.

Nesse contexto, devido à predominância de rios intermitentes em seu domínio territorial, característica essa atribuída à natureza do solo e subsolo que não permite um escoamento subterrâneo adequado para manter os cursos d'água fluindo continuamente, a construção de reservatórios artificiais superficiais emergiu como principal método para o armazenamento de grandes quantidades de água no estado e sua transferência no tempo (Campos, 2022).

Na Região Hidrográfica dos Sertões de Crateús (RHSC) o cenário não é menos desafiador. A RHSC é composta por nove municípios e abriga uma população de 263.155 habitantes, conforme apontado pelo Instituto Brasileiro de Geografia e Estatística – IBGE (2022), exemplifica os desafios enfrentados pelo estado no que se refere ao abastecimento de água. Pois, embora a região possua rios de significativa importância para os municípios locais, a qualidade e o curto período do ano em que esses fluem torna indispensável a utilização de reservatórios superficiais artificiais para que esse recurso tenha garantia de disponibilidade (Rosa; Alexandre; Cid, 2021).

Devido à composição geológica do semiárido nordestino que dificulta o reabastecimento do lençol freático, resultando em baixo potencial hidrogeológico, a construção de açudes para reter águas pluviais e, em alguns casos, perenizar rios, é considerada de

importância singular para atender às necessidades de regiões que enfrentam situações de déficit hídrico (Novais; Carvalho Júnior; Oliveira, 2022). A análise da operação de reservatórios do sistema de abastecimento dos municípios da RHSC ilustra claramente este cenário, mostrando que a maioria dos reservatórios está operando abaixo de sua capacidade crítica, refletindo diretamente a fragilidade dos sistemas existentes e a dificuldade em expandir para atender a um aumento nas exigências de água, tornando a limitação hídrica um obstáculo ao desenvolvimento regional (Rosa; Alexandre; Cid, 2021).

Diante do contexto de dificuldades hidrológicas enfrentadas pela RHSC, a construção da Barragem Fronteiras surge como uma intervenção crítica. Segundo a Secretaria de Recursos Hídricos – SRH (2023), sua capacidade projetada é de 488,18 hm³, que permitirá a ampliação em cerca de cinco vezes a capacidade de regularização das vazões de água em relação aos atuais 10 reservatórios monitorados na região. Além disso, a SRH (2023) também destaca o desenvolvimento do Projeto Malha d'Água, que inclui o Sistema Adutor Fronteiras – Sertões de Crateús, utilizando o futuro açude Fronteiras como manancial. Este sistema, planejado para beneficiar 10 sedes municipais e 54 distritos com uma população total de 329.479 habitantes, destaca a relevância da barragem na garantia de suprimento de água tratada na região, com uma vazão planejada de 658,11 l/s, no horizonte do ano de 2045.

Dessa forma, a capacidade de previsão de vazão, como as afluentes ao açude Fronteiras, é um elemento central para a gestão eficiente dos recursos hídricos, na qual as previsões e incertezas devem ser analisadas no funcionamento e planejamento do sistema (Souza Filho; Lall, 2003), gerando assim, entre outros impactos diretos, a capacidade do desenvolvimento socioeconômico da região (Araújo; Dantas Neto; Souza Filho, 2015).

Segundo Araújo *et al.* (2020) a previsão de vazões é um problema multivariado e de grande complexidade, existindo diferentes metodologias para sua modelagem e análise. De acordo com Malfatti, Cardoso e Hamburger (2018), as técnicas utilizadas frequentemente nesses casos são: as regressões múltiplas; os modelos autorregressivos com média móvel (ARMA, ARIMA) e os modelos não-lineares, como Redes Neurais Artificiais (RNAs). Os autores ressaltam ainda o grande desafio de desenvolver estimativas precisas de vazão, considerando a complexa interação entre os múltiplos fatores que influenciam o comportamento hidrológico de uma bacia hidrográfica.

A singularidade das precipitações no NEB pode ser explicada pela interação de múltiplos elementos ligados aos fluxos atmosféricos globais (Silva; Santos, F; Santos, J, 2021), os quais são fortemente influenciados pela variabilidade da Temperatura de Superfície do Mar (TSM), exercendo influência direta sobre a qualidade da estação das chuvas (Rodrigues *et al.*,

2021). Dessa forma, as TSMs constituem variáveis importantes na análise do clima global (Souza Neto; Silva; Almeida, 2021), sendo essencial compreender seu comportamento e variabilidade.

Souza Filho e Lall (2003) investigaram a dependência da vazão em relação às TSMs com diferentes defasagens temporais nas bacias do Estado do Ceará, no Nordeste do Brasil, obtendo correlações significativas de até 0,9 e fazendo previsões com até 18 meses de antecedência. Este estudo se alinha com a crescente tendência no campo do aprendizado de máquina, destacada por pesquisas que enfatizam a importância dessa tecnologia na previsão de processos hidrológicos tanto no Brasil quanto internacionalmente, ressaltando assim a evolução e a eficácia das técnicas de previsão de vazão (Baran; Santos, 2019; Luo *et al.*, 2019; Papacharalampous; Tyralis; Koutsoyiannis, 2018; Santos; Lima; Mediero, 2018).

Adicionalmente, no que tange ao campo de previsão hidrológica, as metodologias que empregam redes neurais artificiais têm ganhado destaque. Tais abordagens são reconhecidas como alternativas viáveis para a estimação de vazões, apresentando resultados promissores, conforme evidenciado em diversos estudos (Tang; Fishwick, 1993; Srinivasan; Liew; Chang, 1994; Dias; Cataldi; Ferreira, 2017; Oliveira *et al.*, 2018).

Neste âmbito, dada a complexidade inerente à determinação das vazões afluentes em reservatórios no NEB, o uso de redes neurais artificiais apresenta-se como uma opção promissora. Essa tecnologia oferece uma abordagem para criar modelos capazes de estimar as vazões com um nível de precisão apropriado, a partir de dados hidrometeorológicos (Araújo; Dantas Neto; Souza Filho, 2015).

Em vista disso, a análise deste trabalho se concentra na avaliação da capacidade de previsão da vazão anual média no reservatório da Barragem de Fronteiras, situado na bacia hidrográfica dos Sertões de Crateús. Para a construção dos modelos de previsão de vazão, foram aplicadas técnicas de aprendizado de máquina para a modelagem, incluindo a rede neural *Multilayer Perceptron* (MLP), e a rede neural recorrente *Long Short Term Memory* (LSTM), bem como modelos de regressão Ridge, Lasso e a Regressão Linear Múltipla (RLM). Esses modelos de previsão foram utilizados para explorar a relação entre os dados de vazões anuais afluentes ao açude Fronteiras e os índices de Temperatura de Superfície do Mar.

1.1 Objetivos

1.1.1 Objetivo geral

Este estudo tem o objetivo principal avaliar a previsibilidade de vazões anuais afluentes ao açude Fronteiras a partir de modelos analíticos baseados em aprendizado de máquina utilizando índices de Temperatura de Superfície do Mar como variáveis explicativas.

1.1.2 Objetivos específicos

Como objetivos específicos tem-se:

- a) Identificar os índices TSM que possuem maior influência nas vazões afluentes;
- b) Gerar modelos de previsão de vazão utilizando redes neurais artificiais do tipo MLP e LSTM;
- c) Gerar modelos de previsão de vazão utilizando as regressões do tipo Lasso, Ridge e Linear Múltipla;
- d) Verificar a acurácia dos diferentes modelos de previsão à medida que as variáveis são progressivamente incluídas, através das métricas de avaliação, correlação de Pearson e Coeficiente de Nash-Sutcliffe;
- e) Definir o modelo preditor de melhor desempenho.

1.2 Roteiro de trabalho

Este trabalho é estruturado em cinco seções principais. Na primeira seção é apresentado o contexto e a importância do estudo das vazões no semiárido do Nordeste brasileiro, juntamente com a delimitação dos objetivos principais e específicos da pesquisa. A segunda seção, dedica-se a explorar estudos anteriores relevantes, fornecendo uma base teórica sólida para o trabalho, abordando as temáticas de índices TSM, aprendizado de máquina, técnicas de seleção de variáveis, redes neurais artificiais e regressões lineares.

Na terceira seção, é esclarecido a metodologia utilizada, em que se descreve em o procedimento de pesquisa adotado, enquanto a quarta seção, apresenta-se os resultados obtidos e as discussões em decorrência dos mesmos. A quinta e última seção, conclusão, resume os principais achados e discute suas implicações, destacando a relevância do estudo.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo possui como objetivo apresentar uma revisão da literatura acerca dos principais índices TSM atuantes na região de estudo, assim como dos métodos necessários para a realização do trabalho, como técnicas de seleção de características (*feature selection*) e por fim uma discussão sobre modelos de Redes Neurais Artificiais e Regressão Linear.

2.1 Índices de Temperatura de Superfície do Mar

De acordo com Uvo e Berndtsson (1996), o NEB é afetado por uma variedade de sistemas atmosféricos. Estes incluem o El Niño-Oscilação Sul (ENOS), a Temperatura da Superfície do Mar (TSM), os Ventos Alísios, a Pressão ao Nível do Mar, a Zona de Convergência Intertropical (ZCIT), Frentes Frias, Vórtices Ciclônicos, Linhas de Instabilidade e Complexos Convectivos de Mesoescala.

Adicionalmente, Nobre (2012) foca na região norte do Nordeste, objeto deste estudo, destacando que seu regime de chuvas anual ocorre principalmente de janeiro a abril. Sendo a Zona de Convergência Intertropical (ZCIT) o principal agente dessas chuvas, e sua atividade influenciada pela TSM no Atlântico Equatorial, desempenha um papel significativo na modulação das vazões regionais.

2.1.1 Dipolo do Atlântico

O Dipolo do Atlântico é um fenômeno bifásico, definido pela variação de temperatura entre as águas superficiais do Atlântico Tropical Norte (TNA) e do Atlântico Tropical Sul (TSA). A fase positiva do dipolo é caracterizada por anomalias de temperatura positivas no Atlântico Norte e negativas no Atlântico Sul, enquanto a fase negativa apresenta o padrão inverso (Moura; Shukla, 1981; Nobre; Shukla, 1996). Importante destacar que o dipolo positivo do Atlântico, que representa a diferença entre a anomalia da TSM na Bacia do Oceano Atlântico Norte e no Oceano Atlântico Sul, tende a ser desfavorável às chuvas, influenciando assim o regime pluviométrico das regiões afetadas por esse fenômeno (Ferreira; Mello, 2005).

Portanto, na fase negativa do Dipolo do Atlântico, observa-se um aumento das chuvas acima da média no NEB, enquanto na fase positiva, a região tende a experimentar uma redução das precipitações, ficando abaixo da média (Andreoli; Kayano, 2007). Para descrever as fases do Dipolo do Atlântico, a média da anomalia mensal da temperatura da superfície do

mar é calculada nas regiões 5°N-23,5°N, 15°W-57,5°W (TNA) e 0°-20°S, 10°E-30°W (TSA), conforme indicado pelo *Earth System Research Laboratory* (ESRL) em 2020.

2.1.2 ENOS - El Niño e Oscilação Sul

Segundo Aragão (1998), o El Niño consiste no aquecimento das águas marítimas na região do Pacífico Tropical, estendendo-se da costa do Peru/Equador até o oeste do Pacífico. Por outro lado, a Oscilação Sul é uma reação atmosférica ao El Niño, definida pela variação anômala da pressão atmosférica tropical. A combinação destes dois eventos, El Niño e a Oscilação Sul, dá origem ao fenômeno conhecido como ENOS.

O fenômeno ENOS possui duas fases distintas: a fase negativa, conhecida como El Niño, que se caracteriza pelo aumento da Temperatura da Superfície do Mar representando a fase quente, e a fase positiva, chamada La Niña, marcada pelo resfriamento da TSM, simbolizando a fase fria (Bueno *et al.*, 2020).

Para acompanhar as variações da TSM no Oceano Pacífico e os eventos El Niño/La Niña, foram estabelecidos quatro índices climáticos principais: Niño 1.2, Niño 3, Niño 3.4 e Niño 4. As áreas geográficas correspondentes a cada um desses índices são:

- Niño 1+2, localizado entre 0°–10°S e 90°W–80°W;
- Niño 3, situado entre 5°N–5°S e 150°W–90°W;
- Niño 3.4, abrangendo a região de 5°N–5°S e 170°W–120°W;
- Niño 4, que cobre a área de 5°N–5°S e 160°E–150°W.

2.1.3 Oscilação Decadal do Pacífico

Semelhante ao ENOS, este fenômeno também é dividido em duas fases, sendo a fase positiva quando existe um aumento das temperaturas do pacífico e fase negativa, referente a diminuição das temperaturas (Mantua *et al.*, 1997). Durante a fase positiva do PDO (*Pacific Decadal Oscillation*), observa-se uma diminuição na frequência e intensidade do fenômeno La Niña, enquanto há um aumento na probabilidade de ocorrência do El Niño, que tende a ser mais intenso.

Em contraste, na fase negativa do PDO, há uma maior incidência de eventos La Niña, geralmente com maior intensidade, e uma redução na ocorrência do El Niño, que quando acontece, tende a ser breve e menos intenso (Andreoli; Kayano, 2004). Esta fase negativa é

considerada benéfica para o NEB, pois favorece a formação de chuvas na região (Silva; Galvêncio, 2011).

2.1.4 Oscilação Multidecadal do Atlântico

A Oscilação Multidecadal do Atlântico (AMO) é caracterizada por uma sequência contínua e de longo prazo de variações de temperatura na superfície do mar do Oceano Atlântico Norte. Semelhante à PDO, a AMO alterna entre duas fases distintas – uma fria e outra quente – com cada ciclo durando aproximadamente 20 a 40 anos (Silva, 2013).

As fases quentes e frias da AMO têm um papel crucial na movimentação da Zona de Convergência Intertropical (ZCIT) em uma escala multidecadal. Como resultado, as precipitações nas áreas do Atlântico influenciadas pela ZCIT variam de acordo com essas fases. No NEB, a fase fria da AMO está associada a condições mais úmidas, enquanto a fase quente tende a trazer condições mais secas (Knight; Folland; Scaife, 2006).

2.1.5 Oscilação do Atlântico Norte

A principal forma de variabilidade na circulação atmosférica do Atlântico Norte é a Oscilação do Atlântico Norte (NAO - *North Atlantic Oscillation*), como descrito por Trigo, Osborn e Corte-Real (2002). Este fenômeno é caracterizado pela diferença de pressão entre baixas pressões subpolares e altas pressões subtropicais, notável na variação de pressão entre Stykkishoumur, Islândia, e Ponta Delgada, Portugal. O NAO apresenta um padrão de dipolo norte-sul com anomalias, tendo um centro sobre a Islândia e outro oposto nas latitudes centrais do Atlântico Norte, entre 35°N e 40°N (Oliveira; Cunha; Paiva, 2017).

Marengo e Valverde (2007) observaram que a variabilidade do NAO, juntamente com outros fatores, pode esclarecer as flutuações climáticas no NEB em uma escala de tempo interdecadal.

2.2 Aprendizado de máquina (*Machine Learning*)

Aprendizado de máquina (AM) é um ramo essencial da inteligência artificial, focado no treinamento de algoritmos que são programados para identificar ou reproduzir objetivos específicos por meio do reconhecimento de padrões (Bishop, 2006). A área busca desenvolver programas de computador que possuem a capacidade de aprender a realizar tarefas

com base em sua própria experiência, refletindo uma autonomia e adaptabilidade notáveis desses sistemas (Faceli *et al.*, 2011).

Em consonância com esses avanços, a metodologia AM explora métodos computacionais para adquirir novos conhecimentos, habilidades e estratégias inovadoras para estruturar o conhecimento preexistente, impulsionando assim o progresso da inteligência artificial (Mitchell, 1997). Além disso, a investigação em técnicas de aprendizado baseado em computador fornece um melhor entendimento dos nossos processos de raciocínio, estabelecendo um paralelo entre as operações de mentes artificiais e humanas (Monard *et al.*, 1997).

No domínio da Aprendizagem de Máquina, as tarefas preditivas são essenciais, abrangendo a classificação e a regressão. A classificação tem como meta alocar dados em categorias fixas, um processo vital para tarefas como determinar funções de proteínas ou classificar documentos e imagens, ilustrado por pesquisas de Meng *et al.*, Xu *et al.*, e Barker *et al.* Em contraste, a regressão se concentra na previsão de variáveis numéricas contínuas, como valores de mercado de ações ou volumes de precipitação, com estudos conduzidos por Macedo *et al.* e Di *et al.* demonstrando sua importância em áreas como economia e ciências atmosféricas.

2.3 Seleção de Características (*Feature Selection*)

A *Feature Selection* é uma prática amplamente utilizada no aprendizado de máquina, que envolve a escolha dos atributos mais relevantes, buscando melhorar substancialmente o desempenho e a precisão dos modelos de análise e previsão de dados. Para Silva (2019), a seleção de atributos pode ser compreendida como uma tarefa de otimização, na qual se realiza uma exploração do espaço de atributos para identificar o subconjunto ótimo, aquele que demonstra o desempenho mais vantajoso.

Além disso, como enfatizado por Lee (2005), essa prática desempenha um papel crucial, permitindo a classificação de atributos com base em critérios de relevância, a redução da dimensionalidade do espaço de busca de atributos e a eliminação de dados ruidosos. Assim, a seleção de atributos é uma ferramenta indispensável para aprimorar a qualidade e eficácia dos modelos de aprendizado de máquina na análise de dados e previsão.

Portanto, a seleção de atributos assume uma significativa relevância em contextos em que as instâncias são caracterizadas por uma extensa lista de atributos, cuja importância na representação do conceito permanece desconhecida (Castro *et al.*, 2004).

2.3.1 *Random Forest*

O *Random Forest* é um classificador não paramétrico concebido por Breiman (2001), que emprega uma metodologia de aprendizado em conjunto. Conforme Sheykhmousa *et al.* (2020, p. 5), "aprendizagem em conjunto é um esquema de aprendizagem de máquina (*machine learning*) para aumentar a precisão através da integração de vários modelos para resolver o mesmo problema". Portanto, ao agrupar vários classificadores, o *Random Forest* visa diminuir a variância e, por essa via, fornecer resultados mais consistentes e confiáveis (Sheykhmousa *et al.*, 2020).

Além disso, uma das características do *Random Forest* é sua capacidade de identificar os atributos mais significativos no desenvolvimento da floresta. Como explicado por James *et al.* (2013), o método gera uma lista dos atributos mais importantes, selecionados com base na importância acumulada de cada atributo nas divisões dos nós das árvores individuais que compõem a floresta. Isso permite um entendimento mais profundo dos fatores que mais influenciam o modelo, contribuindo para a eficácia do *Random Forest* como ferramenta de classificação e análise.

2.3.2 *Seleção baseada em filtro (Filter-based selection)*

Na seleção de variáveis, o processo de filtragem dos atributos acontece como uma etapa de pré-processamento, antes da aplicação do algoritmo do modelo desejado. Dessa forma, Lee (2005) salienta que o propósito principal da filtragem é eliminar atributos que sejam irrelevantes ou redundantes antes do início do processo de aprendizagem.

Esta metodologia, portanto, foca nas características dos subconjuntos de dados, que são independentes do algoritmo a ser aplicado posteriormente. Uma prática comum é o uso de medidas estatísticas para essa finalidade, exemplificada pelo algoritmo *Select K Best*. Este algoritmo opera selecionando K atributos com base em uma função de teste estatístico previamente definida, escolhendo aqueles com as maiores pontuações conforme o critério estabelecido (Pedregosa *et al.*, 2011).

2.3.3 *XGBoost*

O *XGBoost*, abreviação de *Extreme Gradient Boosting*, é uma biblioteca de código aberto que oferece uma implementação altamente eficiente do algoritmo de aumento de

gradiente. *Gradient Boosting* é uma classe com um conjunto de algoritmos de aprendizagem de máquina aplicáveis tanto em tarefas de classificação quanto de regressão (Melo, 2023). Dastile, Celik e Potsane (2020) destacam que o *XGBoost* se destaca de outros algoritmos de gradiente por sua capacidade de construir árvores de decisão em paralelo, em contraste com a abordagem sequencial mais comum, uma característica que contribui significativamente para seu alto desempenho e rapidez no processamento, tornando-o amplamente utilizado.

Zheng, Yuan e Chen (2017) explica que o método de aumento de gradiente desenvolve árvores aprimoradas com o objetivo de calcular as pontuações dos recursos de maneira inteligente. Isso revela a relevância de cada recurso no modelo de treinamento, elevando a pontuação daqueles recursos que são cruciais para a tomada de decisões nas árvores aprimoradas.

2.4 Redes neurais artificiais

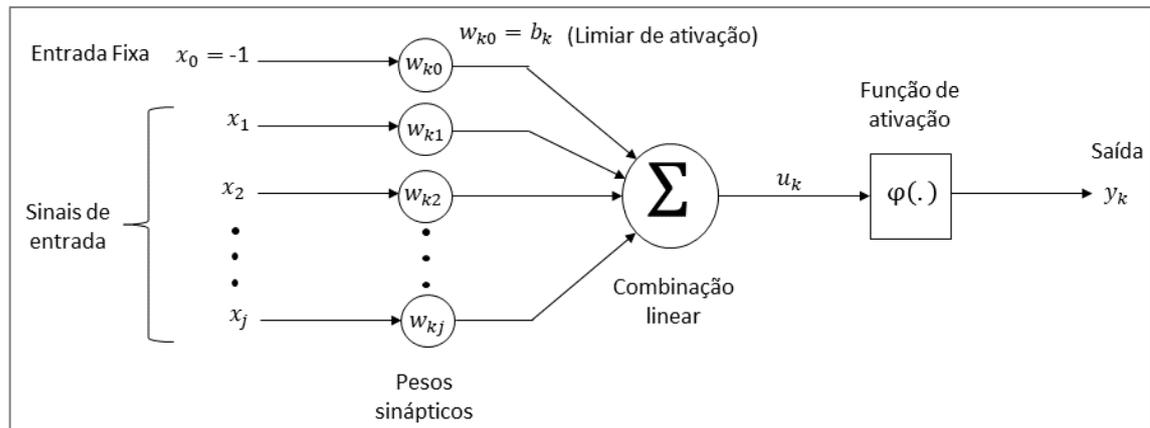
2.4.1 Introdução a redes neurais

As Redes Neurais Artificiais (RNA) são modelos computacionais que emulam a estrutura cerebral humana, visando criar sistemas inteligentes capazes de efetuar tarefas como classificação, reconhecimento de padrões, processamento de imagens e realização de previsões (Haykin, 2001). A rede obtém conhecimento através da análise dos registros de entrada e saída de um determinado processo, utilizando um método de aprendizado que pode ser supervisionado ou independente (Lorenzi *et al.*, 2017).

De acordo com Honorato, Silva e Santos (2017), sob uma perspectiva matemática, as RNAs são equivalentes a uma combinação aditiva de funções previamente estabelecidas, cujos parâmetros, denominados pesos, são ajustáveis. Esta estrutura permite que essas redes representem o comportamento de diversas curvas, abrangendo atributos não-lineares. Desse modo, as RNAs são estruturadas utilizando uma rede de unidades de processamento denominadas neurônios. Além disso, cada neurônio adquire informações de outros através das sinapses, onde essas informações são ponderadas por pesos estipulado pelo desenvolvedor da rede (Vasco, 2020).

A Figura 1 apresenta a estrutura do modelo de um neurônio artificial, em que a informação, ou sinal de entrada, é representada por um vetor $x = [x_1 \ x_2 \ \dots \ x_j]$. Cada componente x_j é ponderada por um peso correspondente w_{kj} , uma componente do vetor de pesos w do neurônio k (Furtado, 2019).

Figura 1 - Estrutura do modelo matemático de um neurônio artificial



Fonte: Santos (2022).

O potencial de ativação do neurônio é o escalar u_k dado pelo produto entre a informação de entrada e os pesos, conforme Equação 1:

$$u_k = \sum_{j=1}^p w_{kj} x_j \quad (1)$$

Esta operação linear combina de forma ponderada a informação de entrada e repassa para a função de ativação $\varphi(u)$ uma operação não-linear que atua sobre o potencial de ativação do neurônio para produzir a saída.

$$y_k = \varphi(u_k) \quad (2)$$

Um outro elemento fundamental para essa estrutura, é o chamado bias ou limiar de ativação, sendo representado por b_k , constitui um peso sináptico associado a uma entrada constante 1, tendo por finalidade promover ajustes com maior liberdade a informação no somatório das entradas. Neste caso a Equação 2 ficaria como descrito na Equação 3:

$$y_k = \varphi(u_k + b_k) \quad (3)$$

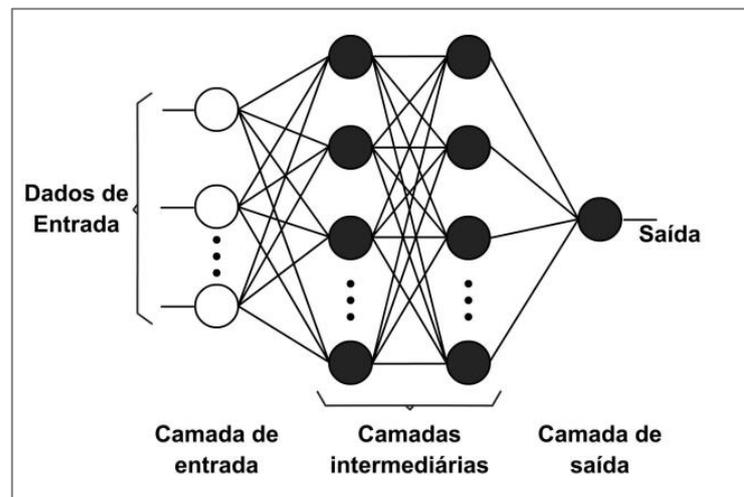
As pesquisas em redes neurais evoluíram das funções de ativação sigmóides, como a sigmoide e a tangente hiperbólica, para a função ReLu, que se destaca pela sua simplicidade e eficácia em algoritmos de otimização e na incorporação de não linearidades na rede (Santos, 2022).

A estrutura dos neurônios em uma rede neural está intrinsecamente relacionada ao algoritmo de aprendizagem usado para o seu treinamento. Essa estrutura, conhecida como arquitetura, descreve a organização dos neurônios nas camadas e os padrões de conexão entre as unidades de processamento (Haykin, 2009).

2.4.2 Rede Neural MultiLayer perceptron

Várias arquiteturas de Redes Neurais Artificiais (RNA) têm sido exploradas, sendo a rede neural de múltiplas camadas amplamente destacada em numerosos estudos acadêmicos (Birikundavy *et al.*, 2002; Dawson *et al.*, 2002; Stokelj *et al.*, 2002; Dawson; Wilby, 2001). Esta arquitetura inclui uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída, o que facilita o desenvolvimento de algoritmos de operação e treinamento. Na prática, como ilustrado na Figura 2, os dados são inseridos pela camada de entrada e processados sequencialmente pelas camadas intermediárias até que o resultado seja obtido e apresentado na camada de saída (Bravo *et al.*, 2008).

Figura 2 - Estrutura de uma Rede Neural MultiLayer Perceptron



Fonte: Florin *et al.* (2011).

Segundo Araújo (2015), o processo de aprendizagem do modelo envolve a modificação de todos os pesos sinápticos (w_{kj}) e limiares (b_k), ajustando-os com base no conhecimento adquirido sobre o fenômeno em estudo. Esta experiência é caracterizada por um conjunto de dados que inclui as entradas e saídas já conhecidas para o problema que está sendo modelado, conforme ilustrado na Equação 4, de acordo com Haykin (2001).

$$\prod_{l=1}^L (\{x_l\}, \{d_l\}) \quad (4)$$

Neste contexto, Π , L , $\{x_l\}$ e $\{d_l\}$ representam, respectivamente, o conjunto de dados de entrada usados na modelagem, a quantidade de entradas e saídas conhecidas no conjunto de dados de treinamento, o conjunto de entrada para os L -exemplos presentes no conjunto de dados de treinamento, e os conjuntos de saída correspondentes a esses mesmos L -exemplos (Araújo *et al.* 2019).

Para o treinamento da rede, é utilizado o algoritmo conhecido como algoritmo de retropropagação de erros, sendo baseado na regra de aprendizagem por correção de erros. Onde a seguir é exposto a demonstração que envolve o algoritmo de retropropagação de erros e a regra do delta generaliza descrita por Haykin (2001). Iniciando pelo sinal do erro do neurônio k (Equação 5), que é obtido pela diferença entre o valor de saída i -ésimo exemplo conjunto de treinamento e o resultado do neurônio $y_k(i)$.

$$e_k(i) = d_k(i) - y_k(i) \quad (5)$$

Sendo $e_k(i)$ e (i) , respectivamente, o sinal de erro do neurônio k e a saída desejada para o mesmo neurônio. O sinal de erro $e_k(i)$ desencadeia um mecanismo de controle que visa realizar ajustes corretivos nos pesos sinápticos dos neurônios k . Estes ajustes são feitos com o objetivo de, gradualmente, alinhar o sinal de saída $y_k(n)$ do neurônio com a resposta desejada $d_k(n)$. Este objetivo é atingido pela minimização da função de custo ou índice de desempenho $E(n)$. A avaliação do desempenho da rede MLP no conjunto de treinamento é realizada utilizando-se o erro quadrático instantâneo, conforme expresso na Equação 6 (Araújo, 2015).

$$E(n) = \frac{1}{2} \sum e_k^2(i) \quad (6)$$

Em que o termo c representa o conjunto que inclui todos os neurônios da camada de saída para o exemplo i do conjunto de treinamento.

No contexto do treinamento em lote, onde o conjunto de pesos sinápticos permanece constante para todos os L exemplos no conjunto de treinamento, a função de custo é definida como o erro quadrático médio $E_{méd}(n)$. Este é calculado conforme a Equação 7.

$$E_{méd}(n) = \frac{1}{2d} \sum_{i=1}^L \sum_{k \in c} [d_k(i) - y_k(i)]^2 \quad (7)$$

Neste caso, n representa o passo no tempo discreto, que corresponde a cada atualização realizada no conjunto de pesos sinápticos durante o processo de treinamento.

Araújo (2015) ainda destaca que o propósito do processo de treinamento é modificar o vetor de pesos sinápticos até que se alcance um valor mínimo para o erro quadrático médio. Portanto, os ajustes contínuos no vetor de pesos sinápticos de um neurônio específico da camada de saída, como o neurônio $y_k(n)$, devem ser feitos seguindo a Equação 8.

$$\{w_{kj}(n + 1)\} = \{w_{kj}(n)\} + \Delta w_{kj}(n) \quad (8)$$

Neste cenário, w_{kj} é o vetor de peso sinápticos na época n e essa modificação é feita de forma iterativa, percorrendo a superfície de $E_{med}(\{w_{kj}\})$ em uma direção que se opõe ao gradiente do erro quadrático médio (∇E_{med}), processo este detalhado na Equação 9. Assim, a rede neural passa por uma série de ajustes precisos, buscando otimizar seu desempenho ao minimizar a função de custo definida.

$$\nabla E_{med}(n) = \frac{\partial E_{med}(n)}{\partial w_{kj}(n)} \quad (9)$$

Dessa maneira, o ajuste do vetor de pesos a atualização do vetor de pesos sinápticos para o neurônio $y_j(n)$ pode ser realizada conforme estabelecido na Equação 10.

$$\{w_{kj}(n + 1)\} = \{w_{kj}(n) - \eta \nabla E_{med}(n)\} \quad (10)$$

Aqui, η representa a taxa de aprendizagem. Desenvolvendo a fórmula do erro quadrático médio (Equação 8) e aplicando os resultados obtidos pela equação do gradiente do erro quadrático médio (Equação 10), chegamos à Regra do Delta (Equação 11). Esta regra oferece uma fórmula para ajustar o vetor de pesos sinápticos para o neurônio $y_k(i)$ da camada de saída.

$$\{w_{kj}(n + 1)\} = \{w_{kj}(n)\} + \frac{\eta}{L} \sum_{i=1}^L \sum_{k \in c} \delta_j \{y_j(i)\} \quad (11)$$

Nesta situação, $w_{kj}(n + 1)$ indica o vetor de pesos sinápticos entre os nós k (saída) e j (entrada) na iteração $n + 1$, e δ_j é o gradiente local do neurônio. No caso de um neurônio da camada oculta, não é possível calcular diretamente seu gradiente local, pois este emite um sinal que se propaga para a próxima camada, sem gerar um sinal de erro distinto. Portanto, ao aplicar o Algoritmo de Retropropagação do Erro, o gradiente local de um neurônio da camada oculta é determinado pela Equação 12 (Araújo, 2015).

$$\delta_j(i) = -\frac{1}{L} f' \left(v_j(i) \sum_{l=1}^L \sum_{k \in c} \delta_k(i) \{w_{kj}(n)\} \right) \quad (12)$$

No qual, o gradiente local do neurônio $y_j(n)$, ou seja, $\delta_k(i)$, é determinado conforme descrito na Equação 13.

$$\delta_k(i) = e_k f'(v_k(n)) \quad (13)$$

Nesse caso, $v_k(i)$ é o campo local induzido do neurônio $y_k(i)$ no i -ésimo exemplo dentro do conjunto de treinamento.

Finalmente, para melhorar a eficácia da taxa de aprendizagem η e evitar instabilidades, é recomendada a inclusão de uma constante de momento α , como sugerido por Haykin (2001). Com base nisso, a atualização dos pesos sinápticos, conhecida como Regra do Delta Generalizada, é realizada conforme a Equação 14.

$$\{w_{kj}(n+1)\} = \{w_{kj}(n)\} + \frac{\eta}{L} \sum_{i=1}^L \sum_{k \in c} \delta_j(i) \{y_j(n)\} + \alpha \{\Delta w_{kj}(n-1)\} \quad (14)$$

As MLP possuem elevadas possibilidades de aplicações em diversos problemas relacionado as mais diferentes áreas do conhecimento, sendo considerada uma das arquiteturas mais versáteis quanto à aplicabilidade. As áreas com maior potencial de aplicação são: aproximação universal de funções, reconhecimento de padrões, identificação e controle de processos, previsão de séries temporais e otimização de sistemas (Silva; Spatti, Flauzino, 2010).

Pesquisas revelaram que uma única camada intermediária em uma RNA pode simular qualquer função não-linear complexa, sendo o MLP com três camadas suficiente para prever e modelar diversos problemas de recursos hídricos (Santos; Silva, 2014). Além disso, estudos hidrológicos que adotaram a configuração de apenas uma camada intermediária alcançaram resultados satisfatórios na aproximação de funções, conforme demonstrado em várias pesquisas (Machado, 2005; Batista, 2009; Santos; Silva, 2014; Honorato; Silva; Santos, 2017).

2.4.3 Rede Neural LSTM

Outro tipo de rede neural artificial em destaque é a Rede Neural Recorrente (RNR), que representa uma evolução das redes *feedforward*, incorporando realimentações que permitem o processamento de dados sequenciais e as tornam mais adequadas para tarefas com dependências temporais ou espaciais (Gomes, 2005). Isso significa que as RNRs processam entradas de duas origens distintas, uma do presente e outra do passado, conferindo-lhes uma capacidade de memória e tornando-as mais semelhantes ao modo como os humanos processam informações (Zaremba; Sutskever; Vinyals, 2014).

Dentro da mesma categoria de arquiteturas recorrentes, encontram-se as Redes Neurais *Long Short-Term Memory* (LSTM). Essas redes, que mantêm as propriedades fundamentais das redes recorrentes convencionais, distinguem-se pela habilidade de preservar informações por extensos períodos no processamento sequencial, atributo este assegurado por suas células que atuam como unidades de memória, essenciais para a manutenção dos dados ao longo do tempo (Santos, 2019).

As células das Redes Neurais LSTMs têm a função dupla de transportar informações ao longo de toda uma sequência e de discernir quais dados devem ser descartados pela rede em determinadas etapas do processo. A configuração das LSTMs é ilustrada na

Figura 3, enquanto o mecanismo por trás de seu funcionamento é detalhado nas equações subsequentes.

$$f_t = \sigma(w_f \cdot [h_{\{t-1\}}, x_t] + b_f) \quad (15)$$

$$i_t = \sigma(w_i \cdot [h_{\{t-1\}}, x_t] + b_i) \quad (16)$$

$$\tilde{C}_t = \tanh(w_c \cdot [h_{\{t-1\}}, x_t] + b_c) \quad (17)$$

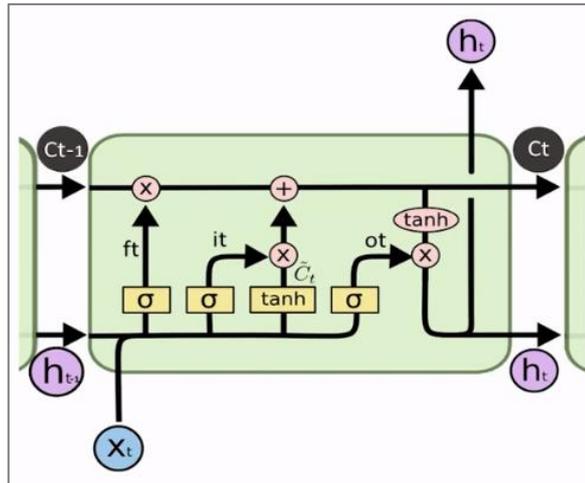
$$C_t = f_t \cdot C_{\{t-1\}} + i_t \cdot \tilde{C}_t \quad (18)$$

$$o_t = \sigma(w_o \cdot [h_{\{t-1\}}, x_t] + b_o) \quad (19)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (20)$$

Na rede no instante t , o x_t representa o input enquanto o h_t é o output da célula. A função σ é a função de ativação logística, e C_t indica o estado atual da célula. O \tilde{C}_t é o candidato a novo estado da célula para o momento t . As redes LSTM são equipadas com três tipos de *gates*: o *gate* de esquecimento f_t , o *gate* de entrada i_t , e o *gate* de saída o_t , que ajudam a regular o fluxo de informações. Os pesos w_f , w_i , w_o , e w_c correspondem aos *gates* de esquecimento, entrada, saída e estado da célula, respectivamente, enquanto os limiares b_f , b_i , b_o , e b_c determinam quando os *gates* devem ser ativados ou desativados na célula (Greff *et al.*, 2016).

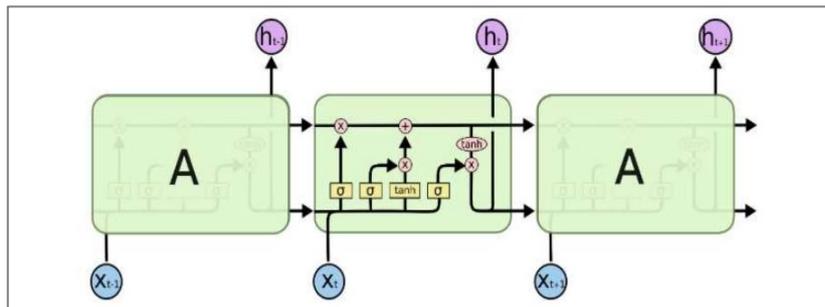
Figura 3 - Representação de uma célula LSTM.



Fonte: Olah (2015).

O *gate* de entrada em uma célula LSTM decide se a informação de entrada deve ser memorizada, enquanto o gate de saída determina se a informação será retida ou descartada no momento t . Esses dados de saída processados e o estado da célula são preservados através deste design, permitindo que as LSTMs aprendam dependências de longo prazo em dados com sequências temporais (Santos, 2019). A Figura 4 demonstra a estrutura temporal das redes LSTM.

Figura 4 – LSTM replicada ao longo do tempo.



Fonte: Olah (2015).

2.4.4 Ajuste de hiperparâmetros

O processo de treinamento, ou aprendizado, em Redes Neurais Artificiais envolve a determinação dos pesos da rede por meio da alimentação de um subconjunto dos dados disponíveis, buscando alcançar uma saída que represente a melhor aproximação da função alvo e aprimorando o desempenho da rede de acordo com um critério definido (Araújo *et al*, 2022; Santos; Silva, 2014).

A avaliação do desempenho durante o treinamento de uma rede neural exige atenção especial à configuração do número de neurônios em cada camada. Braga, Carvalho e Ludermir (2007) argumentam que uma maior quantidade de neurônios incrementa a complexidade da rede e expande o leque de soluções que ela pode aprender. No entanto, uma configuração com muitos neurônios pode levar a um problema de sobreajuste, conhecido como *overfitting*, onde a rede apresenta baixo erro de calibração e elevado erro de previsão. Para prevenir o *overfitting*, é essencial otimizar o número de neurônios na camada intermediária, o que pode ser feito ajustando-se o número de neurônios para alcançar os resultados desejados (Cerqueira *et al.*, 2001; Braga; Carvalho; Ludermir, 2007).

Nesse sentido, o dropout é um mecanismo projetado para prevenir o sobreajuste em modelos de aprendizado de máquina. Funciona ao escolher aleatoriamente um conjunto de neurônios em cada camada durante o treinamento e anulando sua saída, isto é, ajustando o valor para zero. Isso ajuda o modelo a manter a capacidade de generalização para dados não vistos (Srivastava *et al.*, 2014).

Os otimizadores são ferramentas essenciais em redes neurais, responsáveis por encontrar os valores ideais dos pesos para reduzir o erro de previsão. Alguns dos otimizadores mais comuns incluem Adam, *Stochastic Gradient Descent* (SGD) e RMSProp. Esses métodos ajustam os pesos da rede de forma a minimizar o erro durante o treinamento (Vasco, 2020).

Batch Size, ou tamanho do lote, é o termo usado para descrever o número de exemplos de treinamento usados em uma única iteração do processo de aprendizado. É o volume de amostras que a rede examina antes de atualizar os pesos e passar para a próxima iteração.

O número de épocas é um parâmetro que define quantas vezes o conjunto de dados de treinamento inteiro é passado pela rede neural. É vital selecionar cuidadosamente o número de épocas, pois um número muito baixo pode levar a um ajuste insuficiente, onde a rede não aprende o suficiente dos dados. Em contrapartida, um número excessivamente alto de épocas pode causar sobreajuste, fazendo com que o modelo aprenda demais a partir do conjunto de treinamento, prejudicando a sua capacidade de generalização para novos dados (Vasco, 2020).

2.5 Regressão Linear Múltipla

2.5.1 Regressão Linear Múltipla Tradicional

O uso de um modelo de regressão linear é apropriado quando se busca entender, tanto qualitativa quanto quantitativamente, as relações entre diferentes variáveis. Nesse modelo, a variável dependente, representada por y , é aquela cujo comportamento se espera explicar através da variável x , conhecida como variável explicativa ou independente. Se a variável dependente é influenciada por apenas uma variável independente, trata-se de um modelo de regressão linear simples. Por outro lado, se existem múltiplas variáveis explicativas que influenciam a variável dependente, o modelo é classificado como de regressão linear múltipla (Chein, 2019).

O modelo de uma regressão linear múltipla, com k variáveis independentes pode ser definido pela Equação 21 (Hoffmann, 2016).

$$Y_i = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \dots + \beta_k X_{kj} + u_j, \quad j = 1, \dots, n \quad (21)$$

- Y_i : Valor da variável dependente para a i -ésima observação.
- α : Intercepto do modelo, valor esperado de Y_i quando todas as variáveis independentes são 0.
- $\beta_1, \beta_2, \dots, \beta_k$: Coeficientes das variáveis independentes, representam a mudança esperada em Y_i por unidade de mudança em X_j , com outras variáveis mantidas constantes.
- $X_{1j}, X_{2j}, \dots, X_{kj}$: Variáveis independentes para a j -ésima observação.
- u_j : Termo de erro para a j -ésima observação, influências em Y_i não explicadas pelas variáveis independentes.
- $j = 1, \dots, n$: Índice das observações, a equação é estimada para n diferentes observações.

Em problemas de regressão linear múltipla, busca-se encontrar os melhores coeficientes β , de tal forma a reduzir os erros. Para obter os coeficientes β que melhor se ajustem aos dados da regressão, é realizada a minimização da soma dos quadrados dos erros, essa minimização é definida como sendo o estimador do modelo.

A análise tradicional de regressão linear múltipla pode enfrentar desafios, especialmente em situações de multicolinearidade ou quando os conjuntos de dados possuem alta dimensionalidade (Casagrande, 2016).

2.5.2 Regressão Lasso e Ridge

Para mitigar esse problema de multicolinearidade, uma estratégia eficaz é empregar as técnicas de regressão Ridge e Lasso. Estas técnicas introduzem um pequeno viés no cálculo dos estimadores dos coeficientes de regressão, diferentemente do método convencional, reduzindo assim a singularidade do sistema. Esta abordagem resulta em uma diminuição dos erros quadráticos médios, pois, apesar de introduzir um viés no estimador, reduz significativamente a variância dos mesmos (Casagrande, 2016).

Na abordagem tradicional da regressão linear múltipla, a Equação 22 define o erro quadrático. Este erro é crucial, pois sua minimização é necessária para a obtenção das melhores estimativas dos coeficientes.

$$u_i^2 = \left(\sum_{k=1}^n \left(Y_k - \alpha - \sum_j^n (\beta_j X_{kj}) \right) \right)^2 \quad (22)$$

De forma semelhante ao procedimento na regressão linear tradicional, a determinação dos coeficientes ótimos nas regressões Lasso e Ridge também envolve a minimização da Equação 22. Contudo, nesta abordagem, adiciona-se um termo específico conhecido como fator de penalização, que é intrínseco a cada modelo. As equações para o erro quadrático a ser minimizado nas regressões Lasso e Ridge são representadas, respectivamente, pelas Equações 23 e 24. É importante notar que o parâmetro lambda, que é crucial em ambos os modelos, é determinado através do método de validação cruzada.

$$u_i^2 = \left(\sum_{k=1}^n \left(Y_k - \alpha - \sum_j^n (\beta_j X_{kj}) \right) \right)^2 + \lambda \sum_{j=1}^n \beta_j^2 \quad (23)$$

$$u_i^2 = \left(\sum_{k=1}^n \left(Y_k - \alpha - \sum_j^n (\beta_j X_{kj}) \right) \right)^2 + \lambda \sum_{j=1}^n \beta \quad (24)$$

3 METODOLOGIA

A pesquisa é um processo vital para a construção de conhecimento, focada na geração de novas informações e na validação ou contestação de conhecimentos pré-existentes (Clark; Castro, 2003). Zanella (2011) destaca sua importância em desenvolver modelos confiáveis e precisos, essencial no contexto deste trabalho. A pesquisa aplicada, como Silva e Menezes (2005) descrevem, visa gerar conhecimentos para aplicação prática, direcionados à solução de problemas específicos. Este enfoque é alinhado com a natureza deste estudo, que tem o objetivo de resolver um problema prático: a previsão de vazões afluentes no reservatório em questão.

Quanto à sua classificação, esta pesquisa é explicativa, um tipo que, segundo Gil (1991), busca identificar os fatores determinantes ou contribuintes para a ocorrência de fenômenos. Este enfoque é crucial ao lidar com variáveis como os índices de Temperatura da Superfície do Mar (TSM), fundamentais para entender e prever as variações nas vazões afluentes.

A abordagem metodológica abordada neste estudo é caracterizada como quantitativa, definida por Zanella (2011) como o uso de instrumentos estatísticos na coleta e tratamento de dados, visando medir as relações entre variáveis. Neste estudo, as redes neurais, uma parte integrante desta abordagem, permitem modelar relações complexas e não lineares entre variáveis de foco desta pesquisa. Esta abordagem alinha-se também com um dos objetivos específicos deste estudo, que consiste em avaliar a precisão de diferentes modelos de previsão, utilizando métricas estatísticas específicas.

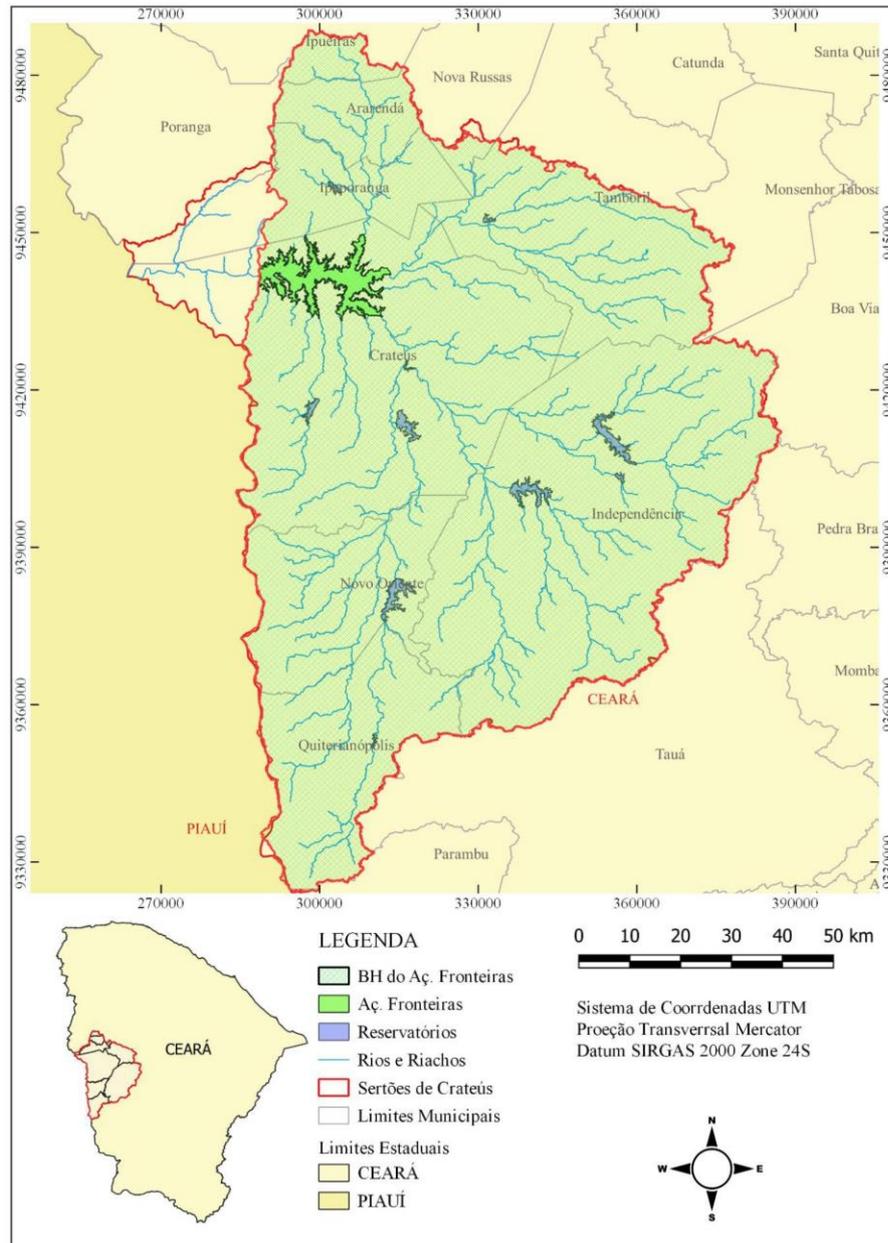
3.1 Região de estudo

A presente pesquisa foca na bacia hidrográfica do açude Fronteiras (Figura 5), uma região de significativa importância hidrológica situada na Região Hidrográfica dos Sertões de Crateús. Esta bacia representa uma área de drenagem de pouco mais de 10.300 km², o que corresponde a aproximadamente 95% da área total da Região Hidrográfica dos Sertões de Crateús. Essa extensão abrange os 10 reservatórios monitorados pela Companhia de Gestão dos Recursos Hídricos (COGERH), conforme indicado pela SRH (2023).

A Região Hidrográfica dos Sertões de Crateús, onde se localiza a bacia do açude Fronteiras, está situada na região oeste do estado do Ceará. Esta área econômica é composta por um total de nove municípios e abrange uma superfície total de 12,82 mil quilômetros

quadrados, conforme relatado pelo Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE, 2017).

Figura 5 – Bacia Hidrográfica do Reservatório Fronteiras.



Fonte: Sousa (2022).

A RHSC apresenta uma geologia complexa, influenciada pela interseção da Bacia Sedimentar do Maranhão-Piauí e do escudo cristalino Pré-Cambriano, e uma geomorfologia com topografias suaves e altitudes entre 200 e 500 metros. Seu clima tropical quente semiárido é marcado por variações na umidade e pluviosidade, influenciado por fatores atmosféricos e geográficos, enquanto a vegetação predominante é a Savana-estépica, adaptada a estas condições. A região também enfrenta altas taxas de evaporação, refletindo em uma umidade

média anual de aproximadamente 59,5%, impactando os reservatórios da área (Secretaria dos Recursos Hídricos, 2021).

3.2 Levantamento dos dados

Os dados das anomalias de temperatura da superfície do mar (TSM) empregados nas análises conduzidas neste estudo foram retirados do portal do *Earth System Research Laboratory* (ESRL), uma divisão do *National Oceanic and Atmospheric Administration* (NOAA¹).

Adicionalmente, anomalias específicas, como NINO3, TSA e TNA, foram extraídas da base de dados de TSM desenvolvida por Kaplan *et al.* (1998). Estes dados podem ser encontrados no *International Research Institute for Climate and Society* (IRI²). Optou-se por esta fonte de dados, uma vez que ela apresenta uma série histórica mais abrangente e está temporalmente alinhada com as outras variáveis consideradas neste estudo.

Para a escolha dos índices TSMs utilizados neste estudo, priorizou-se aqueles que têm sido frequentemente empregados em dissertações e artigos científicos relacionados ao Nordeste brasileiro, com um enfoque particular no estado do Ceará. Este critério de escolha baseou-se na relevância e recorrência da aplicação destes índices em publicações acadêmicas, ilustrado pelos trabalhos de Ferreira e Melo (2005), Araújo *et al.* (2020), Anjos (2015), Araújo Júnior, Souza Filho e Barreto (2019), Fernandes e Silveira (2019). O Quadro 1 a seguir apresenta esses índices.

Quadro 1 - Índices TSM utilizados

Sigla	Nomenclatura (inglês)	Nomenclatura (português)
TAD	<i>Tropical Atlantic Dipole</i>	Dipolo do Atlântico Tropical
TNA	<i>Tropical Northern Atlantic</i>	Atlântico Tropical Norte
TSA	<i>Tropical Southern Atlantic</i>	Atlântico Tropical Sul
NINO1.2	<i>Niño 1+2</i>	Niño 1+2
NINO3.4	<i>Niño 3.4</i>	Niño 3.4
NINO3	<i>Niño 3</i>	Niño 3
NINO4	<i>Niño 4</i>	Niño 4
NAO	<i>North Atlantic Oscillation</i>	Oscilação do Atlântico Norte
BEST	<i>Bivariate ENSO Index</i>	ENSO bivariada ou Índice ENSO "Melhor"
PDO	<i>Pacific Decadal Oscillation</i>	Oscilação Decadal do Pacífico
AMO	<i>Atlantic multidecadal oscillation</i>	Oscilação Multidecadal do Atlântico

Fonte: Elaboração própria (2023).

¹ NOAA – disponível em: <http://www.esrl.noaa.gov/psd/data/climateindices/list>.

² IRI - disponível em: <http://iridl.ldeo.columbia.edu/SOURCES/.KAPLAN/>.

O período analisado na série histórica de vazões vai de 1913 a 2016, sendo que a escolha deste intervalo foi guiada pela disponibilidade da série mais extensa e compatível de dados dos índices TSM, que se estende de 1912 a 2015. A utilização dessas anomalias de TSM do ano anterior é fundamental para a previsão de vazões do ano seguinte, pois nesta pesquisa admite-se essa relação temporal entre os padrões de TSM e as variações climáticas que influenciam diretamente o ciclo hidrológico.

A série de vazões utilizadas nesse estudo tem como base o trabalho de Rosa e Alexandre (2021), que realizaram a calibração dos parâmetros do SMAP mensal a partir dos dados de vazões da estação fluviométrica Fazenda Boa Esperança (Código ANA – 34750000), também localizada no rio Poti, à montante do açude Fronteiras, onde a capacidade de saturação do solo ($SAT = 1736$), parâmetro que controla o escoamento superficial ($PES = 2,7$), coeficiente de recarga do reservatório subterrâneo ($CRec = 0$) e constante de recessão do hidrograma ($K = 3$).

Os dados de vazões afluentes ao açude Fronteiras foram obtidos a partir do estudo realizado por Sousa, Alexandre, Batista (2022). A série de vazões fonte possui 107 anos de dados, entre 1912 e 2018, determinado a partir de modelagem hidrológica chuva-vazão através do modelo SMAP de (Lopes et. al., 1981). A série de precipitação média utilizada para a modelagem chuva-vazão para a bacia hidrográfica do açude Fronteiras foi determinada no âmbito do estudo realizado por Evangelista (2019), a partir de 73 estações pluviométricas da base de dados HidroWeb da Agência Nacional de Águas e Saneamento. Os dados de evaporação mensal utilizados no cálculo da série de vazão pelo SMAP são das Normais Climatológicas entre 1981-2010 da estação Crateús (Código INMET – 82583).

3.3 Estratégia metodológica

Como preditores para a vazão anual média, foram empregados os dados trimestrais das TSM dos seguintes períodos: setembro, outubro e novembro (SON); junho, julho e agosto (JJA); março, abril e maio (MAM); dezembro, janeiro e fevereiro (DJF), refletindo as variações sazonais que podem influenciar o comportamento hidrológico da vazão anual. Dessa forma, os dados TSM empregados para cada trimestre foram definidos como a média dos valores dos índices nos meses que constitui o trimestre. Outrossim, para os valores de vazão anual média, calculou-se a média para os dados mensais de vazão de cada ano.

Dessa maneira, após a coleta e seleção dos dados, foi realizada a classificação das melhores variáveis TSMs, utilizou-se inicialmente os métodos *Random Forest*, *Select K Best* e

XGBoost para avaliar a influência dessas variáveis na vazão. Cada método fornece uma perspectiva da importância das variáveis TSM em relação a vazão. Após essa análise individual, aplicou-se o Método de Copeland (Copeland, 1951) para integrar esses resultados em um único ranking unificado. Esse método compara as variáveis em pares, atribuindo pontos com base no ranqueamento da importância relativa em cada método, resultando em um ranking final que reflete de maneira equilibrada e abrangente a influência das variáveis de TSMS sobre a vazão, o Quadro 2 exemplifica esse ranqueamento das variáveis.

Quadro 2 – Metodologia para o ranqueamento unificado

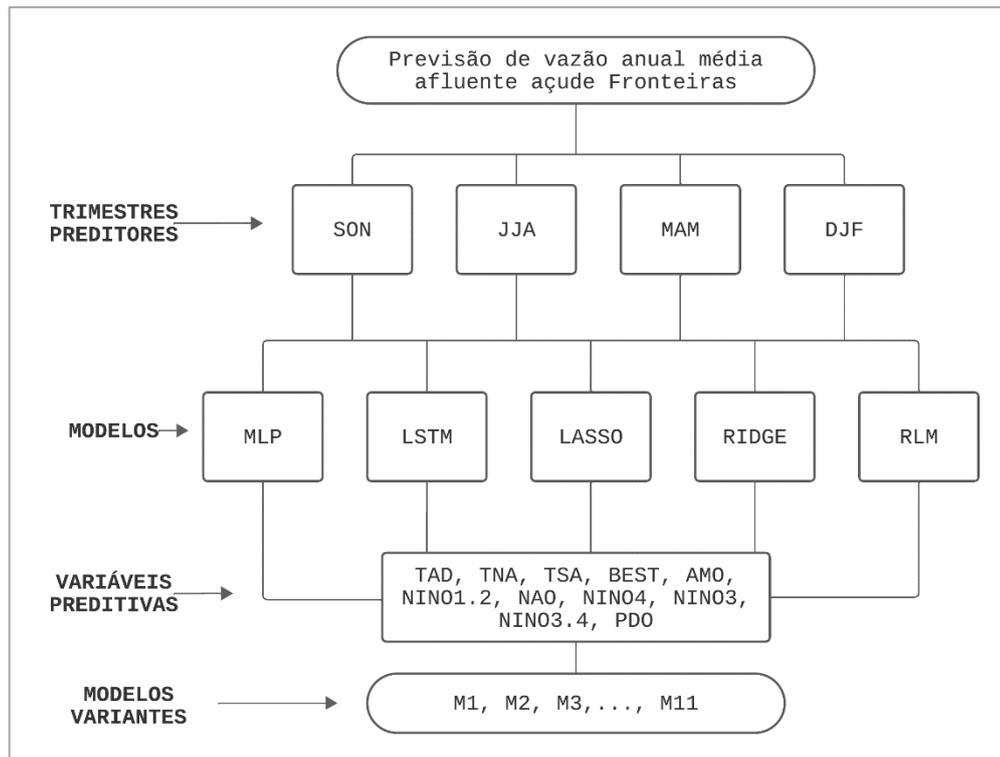
Posição	Ranking 1	Ranking 2	Ranking 3	Ranking unificado Copeland
1°	B	B	C	B
2°	C	A	D	C
3°	A	D	B	D
4°	D	C	A	A

Fonte: Elaboração própria (2023).

Após o ranqueamento das variáveis preditoras, avançou-se para a etapa de previsão de vazão. Para isso, desenvolveu-se cinco modelos de previsão distintos: Rede Neural Multi Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Regressão Lasso, Regressão Ridge e Regressão Linear Múltipla (RLM). Cada um desses modelos de previsão é desenvolvido em 11 variações distintas, correspondentes à inclusão progressiva das variáveis preditoras com base no *ranking* obtido.

O primeiro modelo, denominado M1, utiliza apenas a variável mais bem ranqueada, enquanto os modelos subsequentes, M2 a M11, incorporam gradualmente mais variáveis conforme sua ordem de importância no ranking. Dessa forma, M2 inclui as duas variáveis mais relevantes, M3 as três mais relevantes, e assim por diante, até M11, que utiliza as onze variáveis mais importantes. A Figura 6, apresenta a estratégia para a definição dos modelos de previsão, fazendo a associação entre o trimestre preditor, modelos principais, variáveis preditivas e os modelos variantes. Por exemplo o modelo SON-LSTM-M2, será o modelo associado ao período preditor SON, com o modelo LSTM de variante M2, ou seja, que está usando duas variáveis TSMS, de acordo com o *ranking* unificado do método de Copeland.

Figura 6 – Estratégia para definição dos modelos de previsão.



Fonte: Elaboração própria (2023).

3.4 Classificação por ordem de importância dos índices climáticos

Inicialmente, os dados foram carregados e preparados utilizando a biblioteca Pandas em Python (McKinney *et al.*, 2011). As variáveis independentes (índices TSM) e a variável dependente (Vazão anual média) foram definidas a partir do conjunto de dados, e as técnicas utilizadas para a classificação das variáveis foram desenvolvidas utilizando as bibliotecas do *scikit-learn* (Pedregosa *et al.*, 2011).

3.4.1 *Random Forest*

Inicialmente, para dividir o conjunto de dados em treino e teste, utilizou-se uma proporção de 80% para treino e 20% para teste. Esta divisão foi realizada através de uma amostragem aleatória, criando 50 conjuntos diferentes de dados de treino e teste para validar a robustez do modelo. No contexto da análise com o modelo *Random Forest*, o sistema foi configurado para usar 400 árvores de decisão. O número máximo de variáveis em cada divisão foi limitado a um terço do total de variáveis independentes, visando otimizar a habilidade do modelo em detectar relações não-lineares entre as variáveis, evitando o *overfitting*.

O treinamento do modelo Random Forest foi feito de forma iterativa, utilizando cada um dos 50 conjuntos de treino. A performance do modelo foi então avaliada com base no conjunto de teste correspondente, medindo a acurácia através do erro quadrático médio (MSE).

Para avaliar a importância dos índices climáticos, a métrica '%IncMSE' foi empregada, esta técnica envolve permutar aleatoriamente os valores de cada variável independente e observar o aumento do MSE em comparação ao original. Dessa forma, um incremento significativo no MSE indica uma maior importância da variável para as previsões do modelo. Assim como em Carvalho (2019), a mediana da medida de importância para cada variável foi tomada para classificá-las, possibilitando assim que as anomalias de TSM sejam organizadas por ordem de importância. Para ilustrar visualmente a distribuição e relevância dessas variáveis, foi criado um gráfico de caixa (*boxplot*), facilitando a interpretação da variabilidade e importância dos índices climáticos.

3.4.2 Select K Best

A técnica *Select K Best* foi empregada com a função 'f_regression', ideal para análises de regressão, para avaliar a importância das variáveis independentes em relação à variável dependente, calculando o valor estatístico F para cada uma, e o ajuste do modelo aos dados foi realizado utilizando 'selector.fit(X, y)'.

As pontuações obtidas refletem a relação linear entre cada índice climático e a vazão, em que valores mais altos indicam uma associação mais forte, e sugerem que essas variáveis são preditores significativos da vazão. Para organização e análise, as variáveis foram ordenadas de forma decrescente com base nas pontuações. Com o intuito de proporcionar uma representação visual dos resultados, um gráfico de barras verticais foi gerado, exibindo as variáveis em função de suas pontuações.

3.4.3 XGboost

O modelo *XGBoost* foi configurado com 400 estimadores (número de árvores de decisão) e uma profundidade máxima de 4 para cada árvore, adotando uma fração de 0,33 para a taxa de amostragem por árvore, que define a proporção de recursos considerados na construção de cada árvore. Além disso, um estado aleatório fixo foi estabelecido para assegurar a reprodutibilidade dos resultados.

A relevância de cada índice climático no modelo foi avaliada com base na função de ganho (*gain*), que é intrínseca ao algoritmo *XGBoost*. A função de ganho, neste contexto, mensura o aumento na precisão que cada índice climático traz para as previsões do modelo, ela reflete o quanto a inclusão de uma característica específica ajuda a reduzir a impureza ou incerteza nas previsões. Os índices climáticos foram, portanto, classificados em ordem decrescente de ganho de divisão, o que permite uma avaliação clara de sua importância relativa no modelo. Para visualizar os resultados de forma gráfica, foi criado um gráfico de barras verticais que mostra as variáveis de acordo com suas respectivas pontuações.

3.4.4 Unificação dos Rankings pelo Método de Copeland

O presente estudo aplica o Método de Copeland para a unificação de *rankings*, em que inicialmente, os dados são extraídos de um arquivo, contendo os rankings obtidos pelos métodos anteriores citados. A metodologia adota uma abordagem par a par, contabilizando as vitórias (e empates) e subtraindo as derrotas para cada alternativa, classificando as opções com base nesse resultado (Barba-Romero e Pomerol, 1997). No contexto deste trabalho, cada par de itens é avaliado em cada *ranking*, o item com classificação superior ganha um ponto, enquanto o inferior perde um ponto. Esse processo é repetido para todos os pares em todos os *rankings*, resultando em um escore de Copeland para cada item. Usando o Quadro 1 como exemplo, calculamos a frequência com que cada alternativa ganha e perde, conforme ilustrado no abaixo.

Quadro 3 – Funcionamento método de Copeland

Alternativas	Comparação	Pontuação
a vs. b	6 vs. 9	b ganha 1 ponto
a vs. c	9 vs. 6	a ganha 1 ponto
b vs. c	10 vs. 5	b ganha 1 ponto

Fonte: Adaptado de Amorim (2020).

Após a conclusão do cálculo das pontuações, os itens são ordenados de forma decrescente com base em sua pontuação final de Copeland, estabelecendo assim um *ranking* unificado.

3.5 Modelos de Previsão

Neste trabalho, todos os modelos de previsão foram desenvolvidos utilizando a linguagem de programação Python. Esta escolha foi baseada na eficiência, flexibilidade e no vasto suporte de bibliotecas como *TensorFlow* (Abadi *et al.*, 2016), *Keras* (Chollet, 2015) e

Scikit-learn (Pedregosa *et al.*, 2011), que são essenciais para modelagem, análise e visualização de dados. Além disso, Python se destacou pela facilidade de uso e por sua compatibilidade com ferramentas analíticas.

3.5.1 Pré-Processamento dos dados

Na elaboração dos modelos de previsão deste estudo, que incluem o Perceptron Multicamadas (MLP), e a Rede Neural Long Short-Term Memory (LSTM), e os modelos de regressão Lasso, Ridge e Regressão Linear Múltipla (RLM), foi realizada inicialmente a normalização dos dados, tanto para os dados de entrada, relacionados às anomalias de TSM, quanto para os dados de saída, ligados à série temporal de vazão, esta etapa envolveu a padronização desses conjuntos de dados para uma escala entre 0 e 1. A metodologia utilizada para a normalização é explicada na Equação 25, garantindo uma base de dados uniforme e consistente para a comparação e análise em todos os modelos.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (25)$$

Em que o vetor x consiste nos dados a serem normalizados, e $\max(x)$ e $\min(x)$ representam, respectivamente, os valores máximo e mínimo do conjunto de dados.

3.5.2 Estruturação dos modelos e determinação de hiperparâmetros

Considerando a complexidade inerente à grande quantidade de modelos e hiperparâmetros envolvidos, optou-se por uma seleção inicial baseada em uma série de testes experimentais com os algoritmos de cada modelo. Essa escolha preliminar visou estabelecer um ponto de partida estável, mantendo constantes os hiperparâmetros que consistentemente contribuíram para um desempenho superior nas fases de teste. Tal abordagem metodológica permitiu um foco mais direcionado na otimização e refinamento subsequente dos modelos abordados, dando ênfase a variação dos desempenhos dos modelos em relação ao incremento progressivo das variáveis.

3.5.3 Multi Layer Perceptron

No desenvolvimento deste modelo, é adotada uma abordagem metodológica para a definição e configuração da arquitetura da rede neural. Inicialmente, a construção do modelo é

realizada através da interface de aplicação ‘Sequential’, uma estrutura linear e intuitiva que permite a inclusão sequencial de camadas na rede.

O modelo terá n camadas de entrada de acordo com inclusão de variáveis, seguindo de uma camada oculta (camada densa), que é configurada com oito neurônios. A função de ativação escolhida para esta camada é a ‘ReLU’ (*Rectified Linear Unit*), que permite introduzir não-linearidades no modelo, além disso Gharat (2019) destaca que a função de ativação ‘ReLU’ apresenta uma vantagem computacional significativa, sendo capaz de convergir mais rapidamente em comparação com outras funções de ativação em redes neurais

Em sequência, adicionou-se uma cama de *dropout* de 0,2, visando reduzir o sobreajuste (*overfitting*), já na camada de saída, tem-se um único neurônio, refletindo a natureza do problema de regressão enfrentado, onde o objetivo é prever um valor contínuo. A função de ativação selecionada para esta camada é a ‘sigmoid’, em que a escolha desta função é particularmente pertinente para casos em que a saída desejada do modelo está normalizada, dado que ela restringe a saída a um intervalo entre 0 e 1. Ademais, ela é uma das funções de ativação mais utilizadas, sendo apresentada na Equação 26 (Araújo, 2015).

$$\varphi(v) = \frac{1}{1 + e^{(-x)}} \quad (26)$$

Durante a fase de compilação do modelo, é utilizado o otimizador ‘adam’, aliado com a função de perda (*Loss function*), em que foi adotada o ‘mean_squared_error’, uma escolha convencional para problemas de regressão, pois ela quantifica a diferença entre os valores previstos pelo modelo e os valores reais. Além disso, a mesma métrica é empregada para monitorar o desempenho do modelo durante o treinamento, proporcionando uma avaliação contínua e consistente do erro quadrático médio.

3.5.4 Long Short Term Memory

Para este modelo decidiu-se de adotar os mesmos hiperparâmetros que foram utilizados no modelo MultiLayer Perceptron (MLP), esta escolha objetivou-se em garantir uma comparação equitativa entre os dois modelos de redes neurais, possibilitando uma análise voltada ao comportamento do desempenho em relação a inclusão das variáveis preditivas, observando como cada modelo responde ao conjunto de dados, utilizando o mesmo padrão de

configurações. O Quadro 4 destaca os principais hiperparâmetros e características do modelo mencionado:

Quadro 4 – Hiperparâmetros da rede neural LSTM

Hiperparâmetro	Características e detalhes
Interface de Aplicação	Sequential
Camadas de Entrada	n camadas (variável)
Camada Oculta (Camada Densa)	8 Neurônios, Função de Ativação: ReLU
Camada de Dropout	Taxa: 0,2
Camada de Saída	1 Neurônio, Função de Ativação: Sigmoid
Otimizador	Adam
Função de Perda	Mean Squared Error (MSE)
Métrica de Desempenho	Mean Squared Error (MSE)

Fonte: Autoria própria.

3.5.5 Regressões Lasso, Ridge e Linear Múltipla

Para estes modelos de previsão utilizou-se inicialmente a Regressão Linear Múltipla tradicional e depois implementou-se os modelos de regressão com uma abordagem de regularização. A diferença entre RLM tradicional para os modelos Ridge e Lasso destaca-se na estimação do vetor de parâmetros do modelo, pois enquanto a RLM tradicional foca exclusivamente na minimização da soma dos quadrados dos resíduos (RSS), os modelos Ridge e Lasso complementam esse processo com um termo adicional de regularização, como se nota pela Equação 27 (Marinho; Neto; Rocha, 2022).

$$RSS = \sum_{i=1}^n [y_i - (w \cdot x_i + b)]^2 + \alpha \sum_{j=1}^p w_j^2 \quad (27)$$

O ajuste do hiperparâmetro α no modelo de regressão Ridge, foi realizado através da técnica ‘RidgeCV’ do *Scikit-learn* com validação cruzada de 5 folds. este parâmetro atua como um fator de penalização que modula o grau de complexidade e variabilidade na estimativa dos coeficientes do vetor β . Conforme a magnitude de α aumenta, há uma diminuição proporcional na magnitude dos coeficientes de β resultantes do modelo. É importante salientar que, embora essa técnica restrinja os valores dos coeficientes estimados, ela não os reduz a zero, o que significa que não exclui nenhum preditor do modelo, preservando todos na análise final.

O procedimento de calibração do hiperparâmetro α para o modelo Lasso ocorreu de forma semelhante, porém, utilizando a técnica ‘LassoCV’, mantendo a validação cruzada de 5 folds para avaliação e seleção do melhor valor. Neste caso à medida que o valor de α aumenta,

os coeficientes do vetor β tendem a diminuir, com a possibilidade de alguns deles serem reduzidos a zero, o que resulta na exclusão de preditores menos significativos do modelo, diferenciando-o do ‘RidgeCV’, que mantém todos os preditores no modelo.

3.5.6 *Treinamento e validação*

Para todos os modelos descritos, adotou-se o método de validação cruzada, em que a implementação foi realizada usando a classe ‘TimeSeriesSplit’, adequada para dados de séries temporais devido à sua capacidade de preservar a sequencialidade temporal dos dados. Esta técnica divide os dados em conjuntos de treino e teste (validação) de forma sequencial, garantindo que cada teste seja realizado em dados não utilizados durante o treinamento.

No método ‘TimeSeriesSplit’ adotou-se o parâmetro ‘n_splits = 2’ devido não ter uma quantidade de dados tão grande, com essa configuração, o conjunto de dados é, na verdade, dividido em três partes sequenciais. Na primeira iteração da validação cruzada, a primeira parte (primeiros 1/3 dos dados), é usada para treinar o modelo e a segunda parte (próximos 1/3 dos dados) para testá-lo, deixando a terceira parte ainda não utilizada. Na segunda iteração, o modelo é treinado combinando as duas primeiras partes dos dados (primeiros 2/3 dos dados), e a terceira parte é usada para teste (1/3 final dos dados).

Em relação a etapa de treinamento, para o modelo MLP, o treinamento é realizado por meio de 500 épocas com um lote (*batches*) de 10 amostras. Esta configuração permite que o modelo aprenda progressivamente, ajustando seus pesos internos para minimizar a função de perda ao longo do tempo. Já para a rede LSTM, o número de épocas foi reduzido para 200, e o tamanho do lote de treinamento foi ajustado para 5, pois com um número maior de épocas o modelo resultava em resultados piores. Isso é evidenciado no estudo de Siami-Namini e Namin (2018) no qual observou-se que o aumento do número de épocas não necessariamente leva a uma melhoria na precisão dos modelos, em alguns casos, ocorrendo uma deterioração no desempenho, sugerindo o fenômeno do sobreajuste.

3.5.7 *Métricas de avaliação*

Para a análise da performance dos modelos são aplicados inicialmente o Coeficiente de Correlação de Pearson (R), que é a medida da associação linear entre as observações e previsões, sendo definido na Equação 28.

$$R = \frac{1}{N-1} \sum_{i=1}^N \left(\frac{(P_i - \bar{P})}{S_p} \right) \left(\frac{(O_i - \bar{O})}{S_o} \right) \quad (28)$$

Em que O e P são os valores observados e os da previsão, S é o desvio padrão, \bar{O} e \bar{P} são as médias das previsões e observações e N o número de observações. Este indicador não possui dimensão e seus valores podem oscilar entre 1 e -1, em o valor de 1 é indicativo de uma previsão perfeita (Guarnieri, 2006).

A outra métrica de avaliação utilizada para analisar o desempenho dos modelos é o coeficiente de eficiência do modelo de Nash – Sutcliffe (1970), abreviado como NSE, sendo exposto na Equação 29.

$$NSE = 1 = \frac{\sum(Q_{cal}^t + Q_{obs}^t)^2}{\sum(Q_{obs}^t + \bar{Q}_{obs})^2} \quad (29)$$

Neste contexto, Q_{obs} representa a precipitação observada, enquanto Q_{calc} denota a vazão estimada pelos modelos, e t está relacionado à vazão em um dado instante. O índice em questão, que pode oscilar entre $-\infty$ e 1, $NSE = 1$ significa que há uma combinação perfeita entre as medidas de vazão observadas e calculadas. Um valor de NSE de zero sugere que o modelo prediz a vazão com uma acurácia equivalente à média das observações. Valores negativos de NSE indicam que a média de vazão observada é um preditor melhor do que as predições do modelo (Moriassi *et al.*, 2007).

Além disso, com o intuito de avaliar e comparar a eficiência relativa dos modelos preditivos na etapa de validação, adotou-se a fórmula de taxa de variação percentual, tomando como modelo de referência, o modelo de partida M1. Essa medida quantitativa é calculada pela diferença entre a métrica de um modelo-alvo e a métrica de um modelo de referência, seguido pela divisão do resultado pela métrica do modelo de referência. A Equação 30 exemplifica essa medida para a métrica NSE, e o mesmo princípio se aplica à correlação de Pearson, conforme a Equação.

$$V_{NSE}(\%) = \frac{NSE_{Mi} - NSE_{M1}}{NSE_{M1}}, \text{ com } i = 1, 2, \dots, 11. \quad (30)$$

$$V_R(\%) = \frac{R_{Mi} - R_{M1}}{R_{M1}}, \text{ com } i = 1, 2, \dots, 11. \quad (31)$$

4 RESULTADOS E DISCUSSÕES

Nesta seção, serão apresentados os resultados obtidos a partir da análise dos dados relacionados à vazão anual média e sua relação com as variáveis preditivas de Temperatura da Superfície do Mar (TSM).

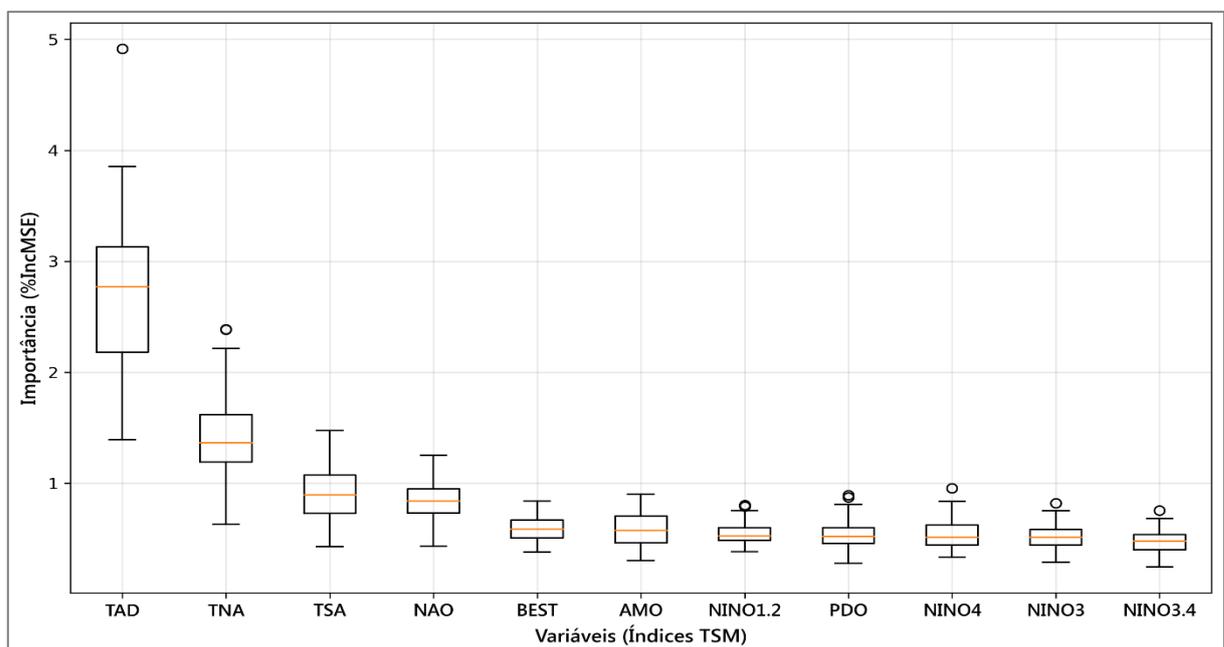
Os resultados são organizados de maneira a refletir a sequência dos períodos preditores ao longo do ano, iniciando com o trimestre SON, seguindo para JJA, MAM e finalizando com DJF. Para cada período preditor, são detalhados os resultados dos métodos analíticos aplicados, os *rankings* derivados e as implicações desses achados. Adicionalmente, são discutidos os resultados dos modelos de previsão de vazão, que foram desenvolvidos em 11 variações distintas, incorporando progressivamente as variáveis predictoras de acordo com a sua importância no ranking unificado obtido pelo método Copeland. Salienta-se que para as análises do *random forest* o símbolo \circ é referente aos *outliers*.

4.1 Análise dos resultados para o período preditor SON

4.1.1 Análise da importância dos índices TSM para o período preditor SON

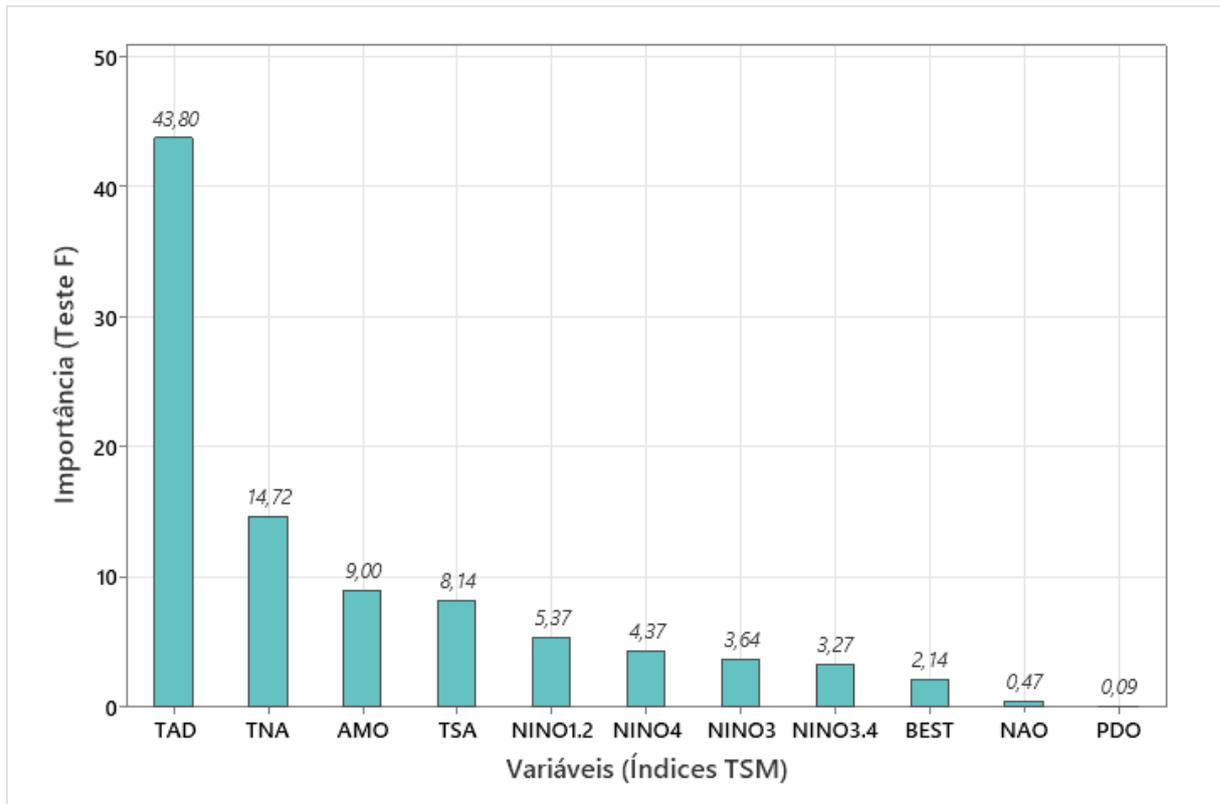
As Figura 7, Figura 8 e Figura 9 apresentam a importância das variáveis para cada uma das técnicas utilizadas para esta avaliação, para o período preditor SON.

Figura 7 - *Boxplot* da importância dos índices TSM no período preditor SON (*Random Forest*)



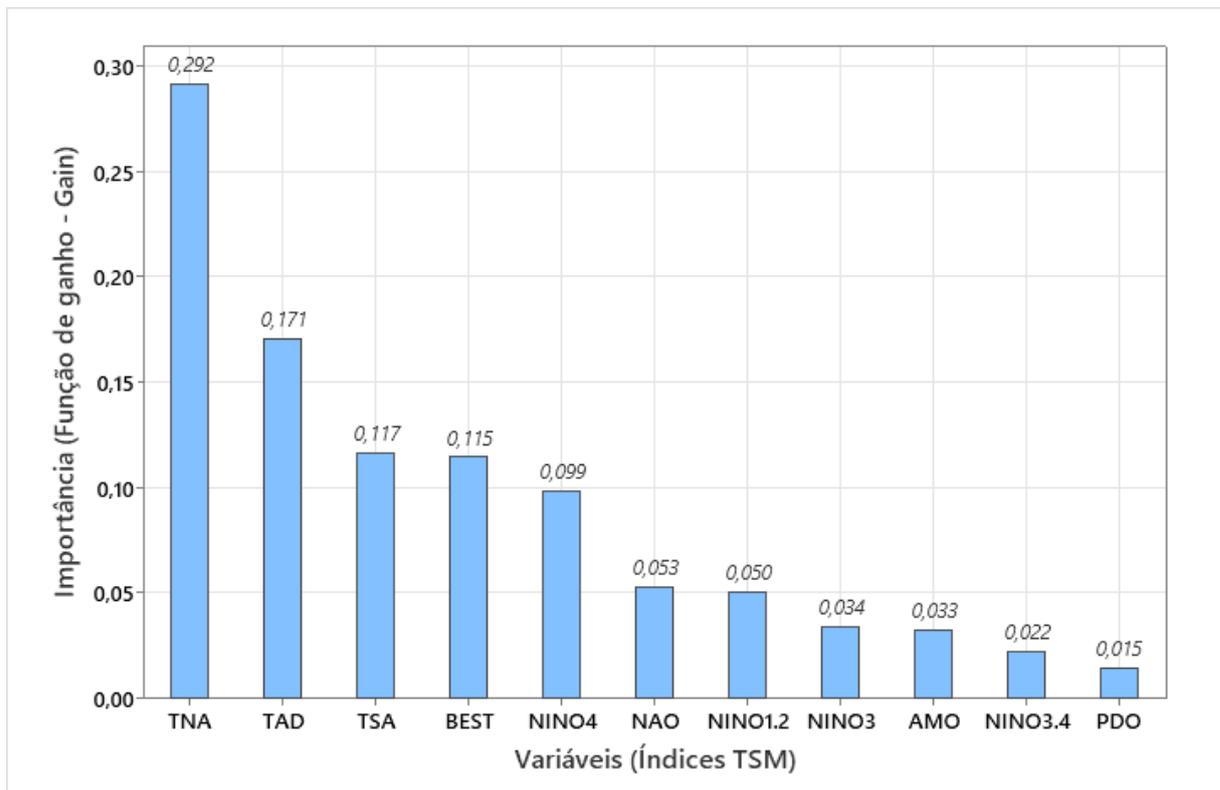
Fonte: Elaboração própria (2023).

Figura 8 - Importância dos índices TSM no período preditor SON (*Select K Best*)



Fonte: Elaboração própria (2023).

Figura 9 - Importância dos índices TSM no período preditor SON (*XGBoost*)



Fonte: Elaboração própria (2023).

O uso do método *Random Forest* para analisar a importância das variáveis de TSM durante o trimestre SON indica uma distribuição variada de importância, como ilustrado pelo gráfico de *boxplot* (Figura 7). O índice TAD mostra-se relevante, contudo, a distribuição dos resultados sugere que outras variáveis também exercem influência significativa. É notável a presença de outliers indicando que algumas variáveis podem ter efeitos extremos em certas condições. A variabilidade das importâncias realça a complexidade das interações entre TSM e vazão, evidenciando uma mediana constante após a variável BEST.

Seguindo para o *Select K Best* (Figura 8), os resultados apontam para o índice TAD como o de maior importância, com uma pontuação de teste F notavelmente alta de 43,80. TNA e AMO também são identificados como preditores significativos, com pontuações de 14,72 e 9,00, respectivamente. Essa abordagem destaca um conjunto de variáveis TSM que são estatisticamente significativas na correlação com a vazão anual média durante o período SON.

Por fim, a análise via *XGBoost* (Figura 9) revela o índice TNA como o mais influente com um ganho de informação (*gain*) de 0,292. Este método apresenta uma perspectiva diferenciada, onde o índice TAD, apesar de ainda importante, não é tão preponderante quanto no *Select K Best*, apresentando um ganho de informação de 0,171. Os índices TSA e BEST também mostram uma importância considerável, com ganhos de 0,117 e 0,115, respectivamente. O *XGBoost* sugere, portanto, uma relevância mais equilibrada entre as variáveis TSM, oferecendo uma visão complementar sobre quais fatores são mais preditivos para a vazão no trimestre SON.

4.1.2 Ranqueamento unificado pelo método Copeland para o Preditor SON

O resultado do Método de Copeland, sintetizou os ranqueamentos dado pelos modelos anteriores, fornecendo uma perspectiva integrada sobre quais índices TSMs são mais preditivos para a vazão anual no período SON. Pois considerou-se a posição relativa de cada índice nos diferentes modelos de classificação. O Quadro 5 apresenta o ranking unificado.

Quadro 5 – *Ranking* Copeland para período SON

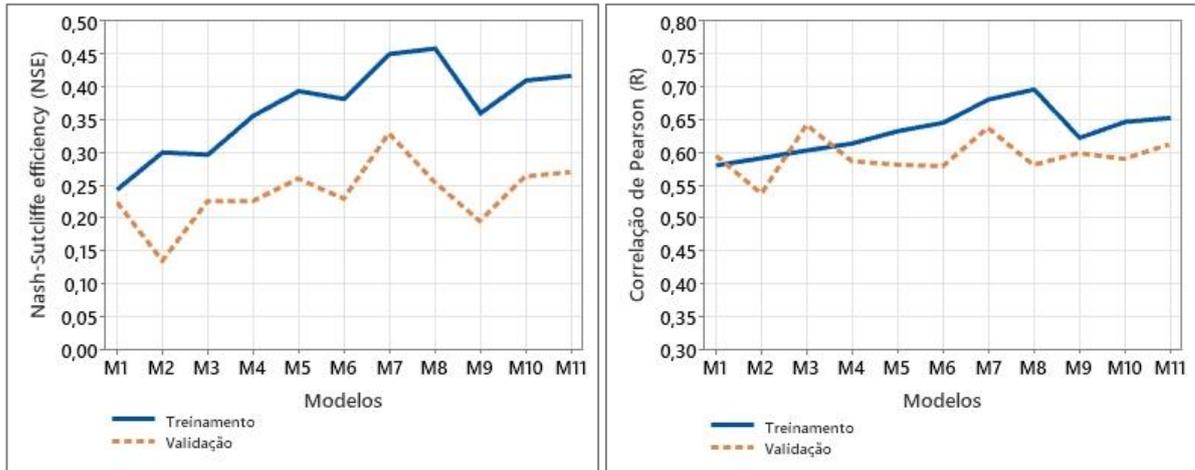
Posição	Variável	Posição	Variável
1°	TAD	7°	NAO
2°	TNA	8°	NINO4
3°	TSA	9°	NINO3
4°	BEST	10°	NINO3.4
5°	AMO	11°	PDO
6°	NINO1.2		

Fonte: Elaboração própria (2023).

4.1.3 Modelos de previsão de vazão – SON

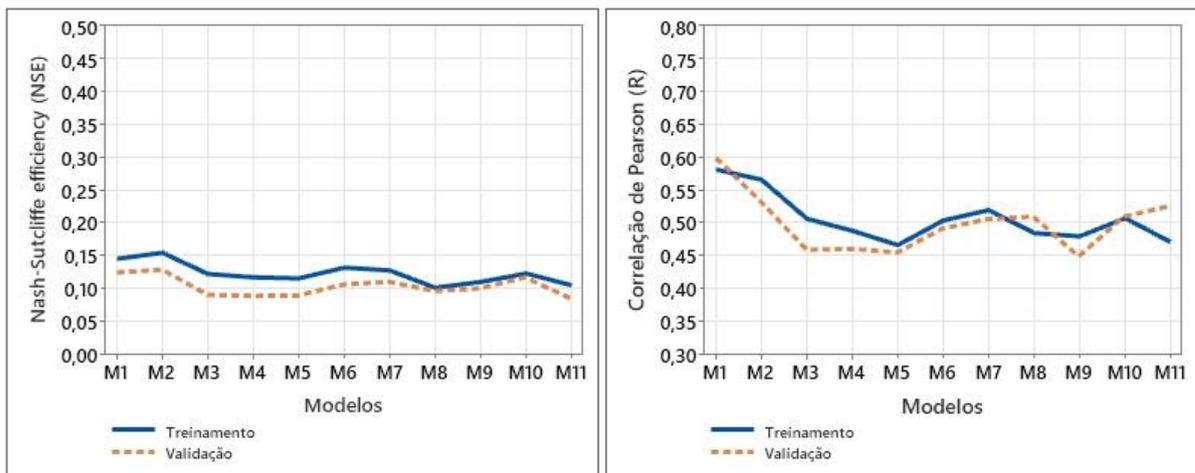
Nas Figuras 10, 11, 12, 13 e 14 apresenta-se a correlação e o NSE entre a vazão predita e a vazão real para cada modelo. Esta análise é realizada ao incorporar os índices climáticos progressivamente seguindo o ranking de Copeland para o preditor SON.

Figura 10 - Métricas de Avaliação do Modelo MLP com Preditor SON (NSE e Correlação)



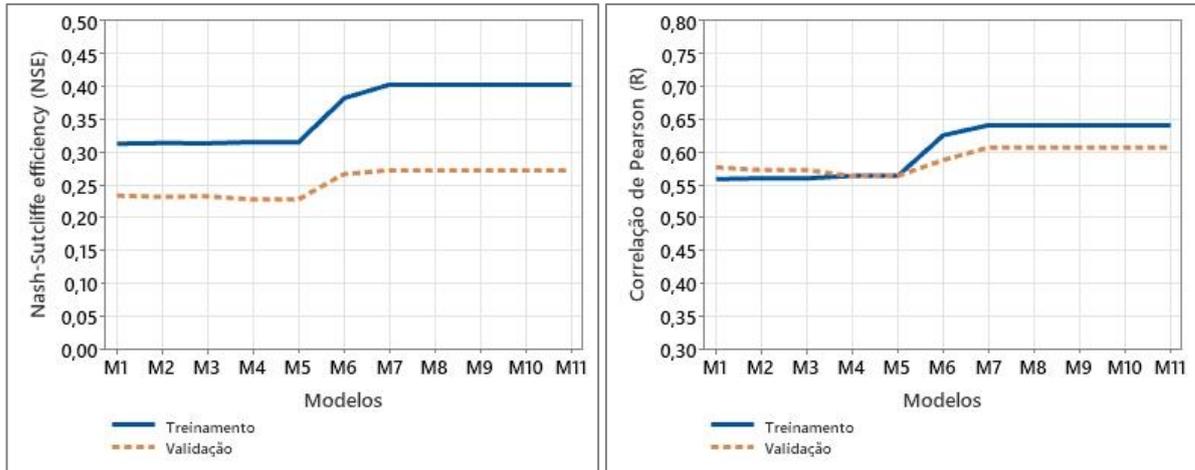
Fonte: Elaboração própria (2023).

Figura 11 - Métricas de Avaliação do Modelo LSTM com Preditor SON (NSE e Correlação)



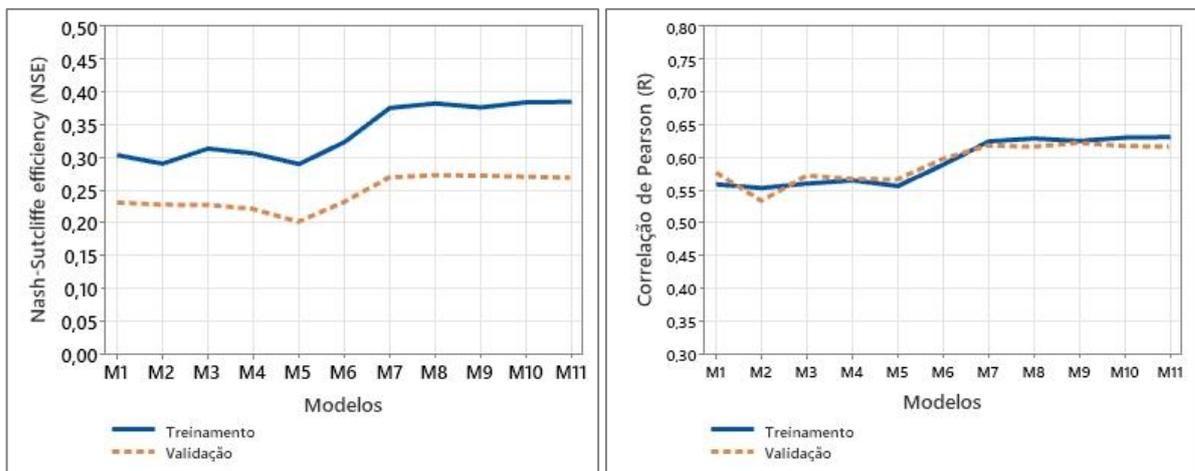
Fonte: Elaboração própria (2023).

Figura 12 - Métricas de Avaliação do Modelo LASSO com Preditor SON (NSE e Correlação)



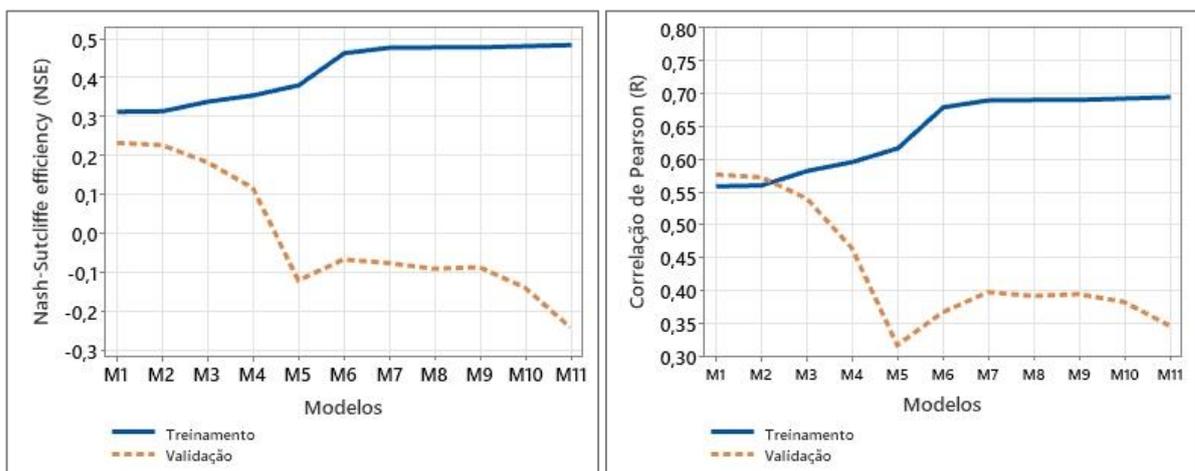
Fonte: Elaboração própria (2023).

Figura 13 - Métricas de Avaliação do Modelo RIDGE com Preditor SON (NSE e Correlação)



Fonte: Elaboração própria (2023).

Figura 14 - Métricas de Avaliação do Modelo RLM com Preditor SON (NSE e Correlação)



Fonte: Elaboração própria (2023).

Analisando a Figura 10, observou-se que o modelo MLP, particularmente, demonstra uma melhoria progressiva nas métricas de desempenho com a adição de previsores, alcançando o pico no modelo M7 (NSE de treinamento de 0,458; NSE de validação de 0,329), indicando que a adição de variáveis até este ponto contribuiu positivamente, acompanhada por uma melhoria consistente no coeficiente de Pearson, indicando um grau considerável de correlação, com R de 0,680 e 0,638 no modelo M7 para as etapas de treinamento e validação, respectivamente. No entanto, a performance variou após M7, sugerindo a possibilidade de sobreajuste com o incremento de mais variáveis.

Por outro lado, os modelos LSTM, em geral, exibem um desempenho mais modesto (Figura 11). Mesmo com a inclusão de variáveis adicionais, não se observa um aumento nas métricas, com o modelo M2 alcançando o maior NSE, sendo de apenas 0,154 no treinamento e 0,127 na validação. Em relação a correlação, observou-se comportamento semelhante, atingindo maior correlação com apenas uma variável (M1), com R de treinamento e validação igual a 0,581 e 0,599, respectivamente; posteriormente apresentando correlações decrescentes ao aumentar o número de variáveis. Esses resultados sugerem limitações no modelo LSTM em capturar a dinâmica dos dados de vazão anual neste contexto específico.

Em contraste, o modelo de regressão Lasso apresentou um salto significativo no desempenho com a inclusão da variável NINO1.2 em M6 com NSE de treinamento igual a 0,382 e NSE de validação igual a 0,266, mantendo uma performance estável de M7 a M11, conforme ilustrado na Figura 12. Isso é evidenciado pelos valores de NSE de treinamento e validação em M7 (NSE de treinamento igual a 0,403; NSE de validação igual a 0,272), que se mantiveram relativamente constantes apesar da adição de mais variáveis, apresentando comportamento semelhante para a correlação. Isso pode indicar que a regularização Lasso impõe uma restrição eficaz na complexidade do modelo, prevenindo o overfitting (sobreajuste) mas também limitando o desempenho máximo alcançável.

A regressão Ridge seguiu uma tendência similar à do Lasso em termos de NSE, com melhorias notáveis no desempenho até M7 (NSE de treinamento igual 0,375; NSE de validação igual a 0,269), após esse ponto a performance se manteve quase uniforme, apresentando pequenas variações, conforme a Figura 13. Da mesma forma, a correlação apresenta perfil semelhante ao NSE, em que os valores R de treinamento e validação se mantêm próximos ao longo de todos os modelos, com destaque de melhoria expressiva do modelo M5 para o M7, saindo de uma correlação de validação de 0,566 para 0,618. Este padrão sugere que ambas as técnicas de regressão regularizada foram eficazes para mitigar o sobreajuste, exibindo métricas de avaliação relativamente estáveis a partir de M7.

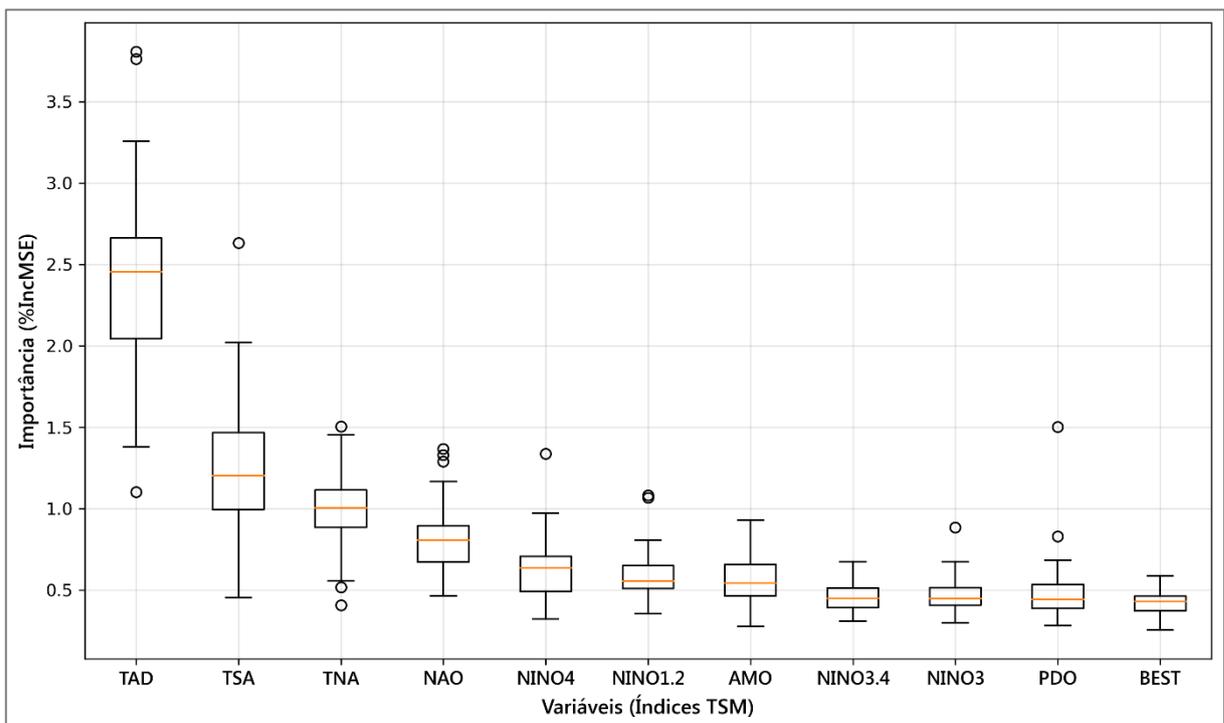
Por fim, o modelo de regressão Linear Múltipla apresentou um problema significativo de sobreajuste, como indicado pelo declínio no NSE de validação a partir de M3, tornando-se negativo em M5 e piorando até M11, atingindo um NSE de treinamento de 0,484 e NSE de teste de -0,243, com as 11 variáveis (Figura 14). Ainda que tenha ocorrido um aumento consistente na correlação de Pearson durante o treinamento até o modelo M11, alcançando um R de 0,694, na fase de validação observou-se uma diminuição correspondente, com o R reduzindo para 0,345, evidenciando a perda de capacidade de generalização do modelo.

4.2 Análise dos resultados para o período preditor JJA

4.2.1 Análise da importância dos índices TSM para o período preditor JJA

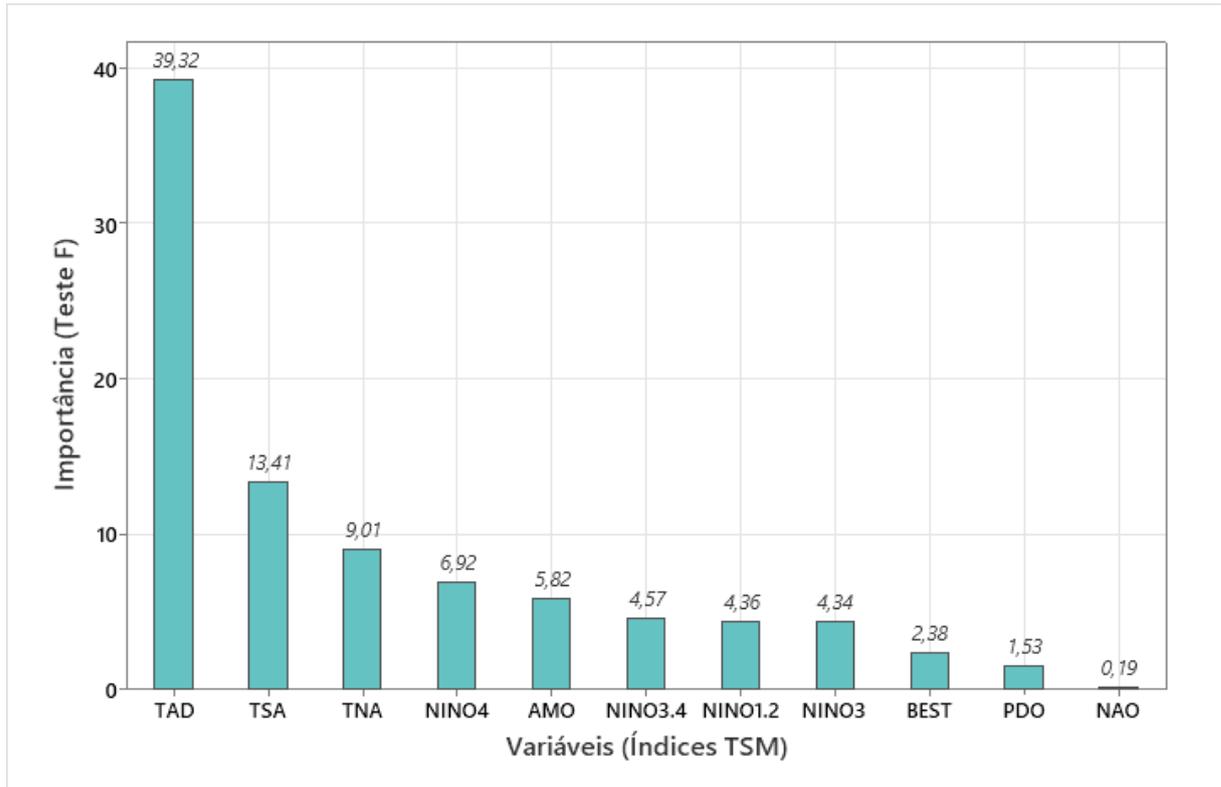
As 15, 16 e 17 apresentam a importância das variáveis para cada uma das técnicas utilizadas para esta avaliação, para o período preditor JJA.

Figura 15 - *Boxplot* da importância dos índices TSM no período preditor JJA (*Random Forest*)



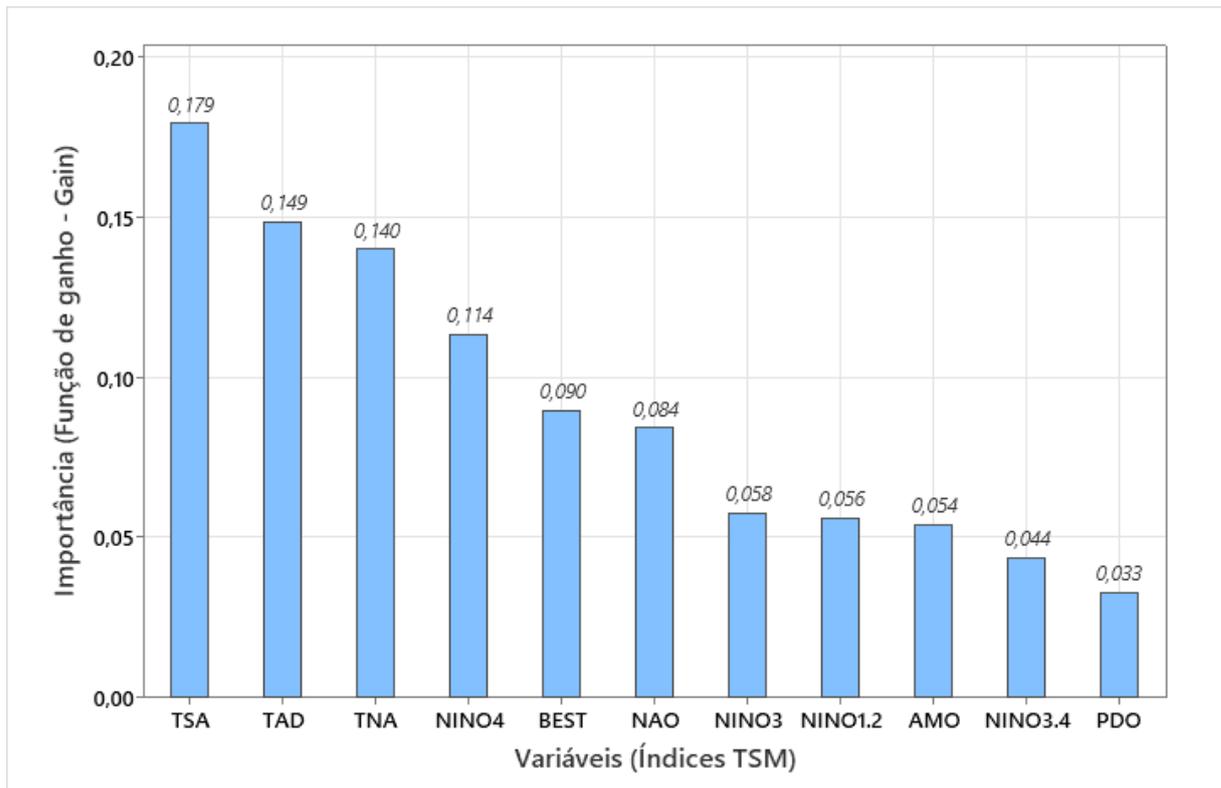
Fonte: Elaboração própria (2023).

Figura 16 - Importância dos índices TSM no período preditor JJA (*Select K Best*)



Fonte: Elaboração própria (2023).

Figura 17 - Importância dos índices TSM no período preditor JJA (*XGBoost*)



Fonte: Elaboração própria (2023).

De acordo com a análise da Figura 15 para o trimestre JJA, o *Random Forest* identificou o índice TAD como o de maior importância, como é evidenciado pelo pico na distribuição da importância relativa no gráfico de *boxplot*. A distribuição ampla e a presença de outliers sugerem que TSA pode ter um impacto variável, mas significativo, na vazão para este preditor. Além do TSA, outros índices como, TNA e NAO também mostram uma importância relevante, indicando que o comportamento da vazão neste trimestre é influenciado por múltiplas variáveis de TSM.

Seguindo a análise da Figura 16, com o *Select K Best*, o índice TAD se sobressai com a maior importância (teste F de 39,32), seguido pelo TSA e TNA. Esta tendência reforça o TAD como uma variável preditora consistente em diferentes métodos. Os resultados destacam um padrão claro de influência destas variáveis TSM, com uma hierarquia distinta de importância, em que as variáveis apresentam importância consideravelmente menor em comparação ao TAD.

Já em relação ao *XGBoost* (Figura 17), este apresenta um perfil ligeiramente diferente para a importância das variáveis TSM. TSA se destaca como a mais importante com um ganho de informação de 0,179, seguido de perto por TAD e TNA. Os índices relacionados ao fenômeno El Niño (NINO4, BEST) também surgem como preditores relevantes neste trimestre, o que pode ser indicativo de padrões climáticos específicos que afetam a vazão para o período preditor JJA.

4.2.2 Ranqueamento unificado pelo método Copeland para o Preditor JJA

Utilizando o Método de Copeland, análogo ao que foi empregado para o período SON, o Quadro 6 exibe um ranking diferenciado para este preditor, evidenciando as variações na influência dos índices TSM na vazão anual para o período preditor JJA.

Quadro 6 – Ranking Copeland para período JJA

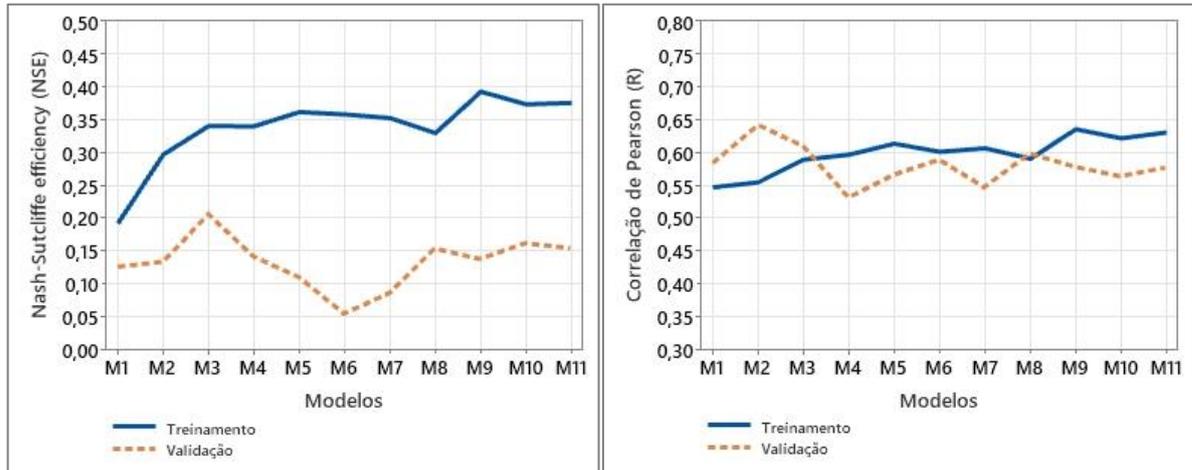
Posição	Variável	Posição	Variável
1°	TAD	7°	AMO
2°	TSA	8°	NINO3.4
3°	TNA	9°	NINO3
4°	NINO4	10°	BEST
5°	NAO	11°	PDO
6°	NINO1.2		

Fonte: Elaboração própria (2023).

4.2.3 Modelos de previsão de vazão – JJA

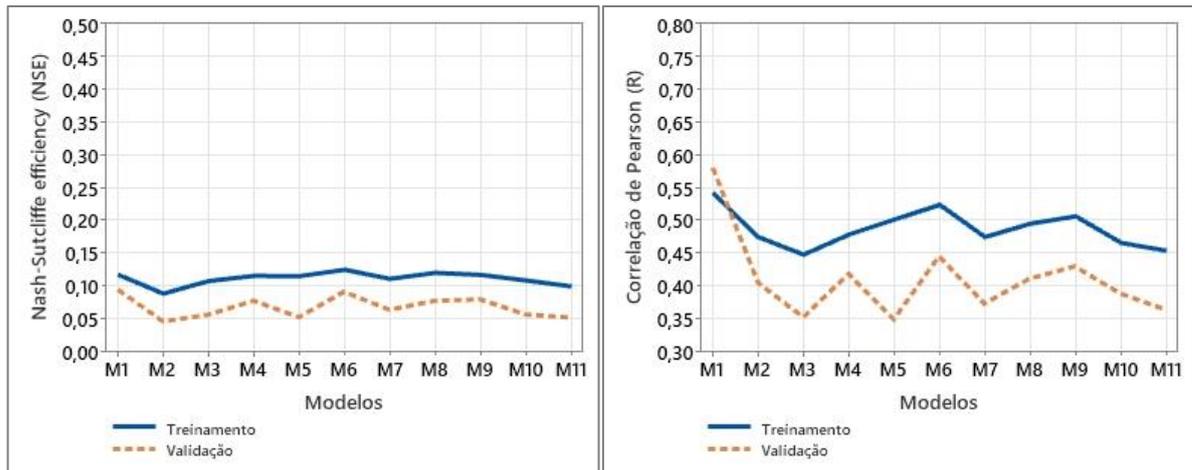
Nas Figuras 18, 19, 20, 21 e 22 apresenta-se a correlação e o NSE entre a vazão predita e a vazão real para cada modelo. Esta análise é realizada ao incorporar os índices climáticos progressivamente seguindo o ranking de Copeland para o preditor JJA.

Figura 18 - Métricas de Avaliação do Modelo MLP com Preditor JJA (NSE e Correlação)



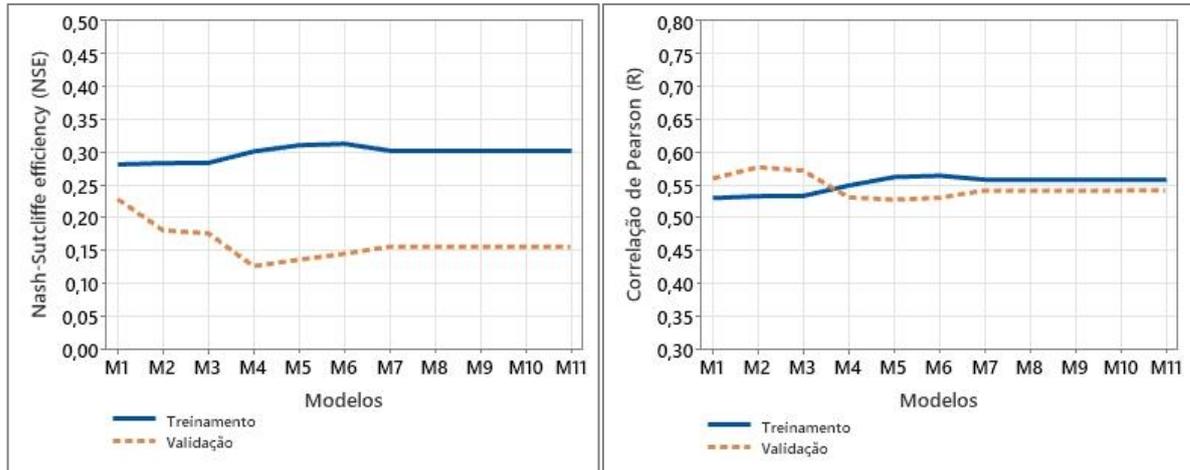
Fonte: Elaboração própria (2023).

Figura 19 - Métricas de Avaliação do Modelo LSTM com Preditor JJA (NSE e Correlação)



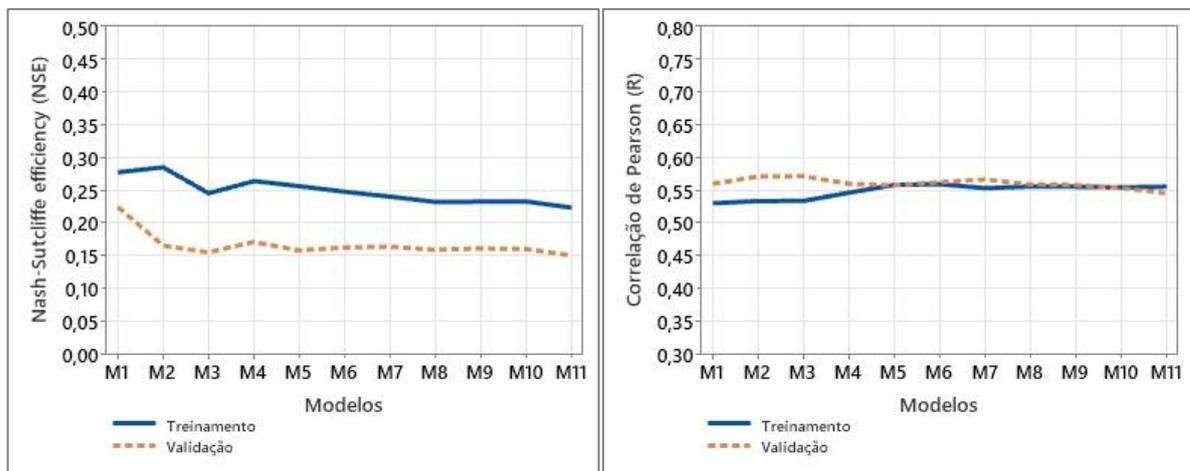
Fonte: Elaboração própria (2023).

Figura 20 - Métricas de Avaliação do Modelo LASSO com Preditor JJA (NSE e Correlação)



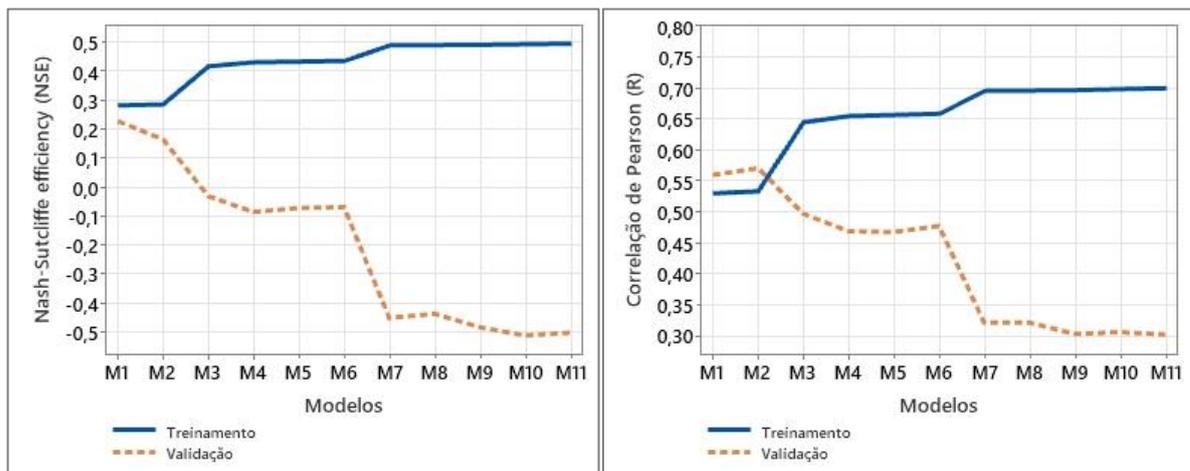
Fonte: Elaboração própria (2023).

Figura 21 - Métricas de Avaliação do Modelo RIDGE com Preditor JJA (NSE e Correlação)



Fonte: Elaboração própria (2023).

Figura 22 - Métricas de Avaliação do Modelo RLM com Preditor JJA (NSE e Correlação)



Fonte: Elaboração própria (2023).

O modelo MLP (Figura 18) mostrou melhoria no NSE de treinamento até M3, passando de 0,191 para 0,340, indicando aprimoramento pela inclusão de variáveis até esse ponto. Após M3, não houve ganhos significativos, com pico no M7 e queda subsequente. Na validação, o NSE mais alto foi em M3 (0,205), caindo em M4 (0,140), sugerindo possível sobreajuste. O sobreajuste ficou mais evidente em M6, com NSE de teste caindo para 0,053, enquanto o de treinamento era alto (0,358). Quanto à correlação de Pearson no treinamento, houve aumento até M9 (0,635), mostrando forte relação linear positiva. No entanto, na fase de validação, a correlação máxima ocorreu em M2 (0,642), e depois declinou, indicando que variáveis adicionais não melhoraram a generalização do modelo para dados nunca vistos.

No Modelo LSTM (Figura 19), a análise do NSE e da correlação de Pearson mostra um padrão menos definido em comparação ao MLP. No treinamento, o NSE inicia em 0,116 em M1 e varia ligeiramente, com um pico em M6 (0,124), indicando melhoria inconsistente com a adição de variáveis. Isso pode ser devido à capacidade do LSTM de lidar com dependências de longo prazo. Na validação, o NSE tem um pico em M6 (0,089), depois declina, sugerindo dificuldades na generalização após adições de variáveis. A correlação de Pearson também atinge o valor mais alto em M6 tanto no treinamento (0,523) quanto na validação (0,444), indicando que a inclusão de variáveis até M6 é benéfica, mas adições posteriores podem não ser tão eficazes.

Em relação ao comportamento da regressão Lasso (Figura 20), o NSE de treinamento começa em 0,281 em M1 e melhora levemente até M6 (0,312), mas estagna a partir de M7, mantendo-se em torno de 0,302. Na fase de validação, o NSE começa em 0,228 em M1, cai até M4 (0,125), aumenta até M6 (0,144) e depois se estabiliza até M11, indicando possível sobreajuste com variáveis adicionais, mas com alguma melhoria ou estabilidade na generalização em modelos posteriores. A correlação de Pearson no treinamento aumenta até M6 e se estabiliza em M7, mantendo-se perto de 0,558 até M11. Na validação, é mais alta em M2 (0,578), diminui em M4 e depois se estabiliza nos modelos seguintes, seguindo um padrão similar ao do NSE.

Para a regressão Ridge (Figura 21), no treinamento, o NSE do modelo começa em 0,277 em M1, tem um leve aumento em M2 (0,285), e depois diminui até M11 (0,223). Este padrão indica que variáveis adicionais após M2 não melhoram a adequação do modelo. Na fase de validação, o NSE inicia alto com o primeiro modelo M1, entretanto nos modelos posteriores apresenta um perfil uniforme, com pouca variação e sem indicação de melhorias. Em relação, a correlação de Pearson, observa-se um perfil pouco variável e uniforme tanto para a etapa de

treino como a de validação, apresentando valores mais altos inicialmente na fase de validação, dando indicativo de sobreajuste neste modelo.

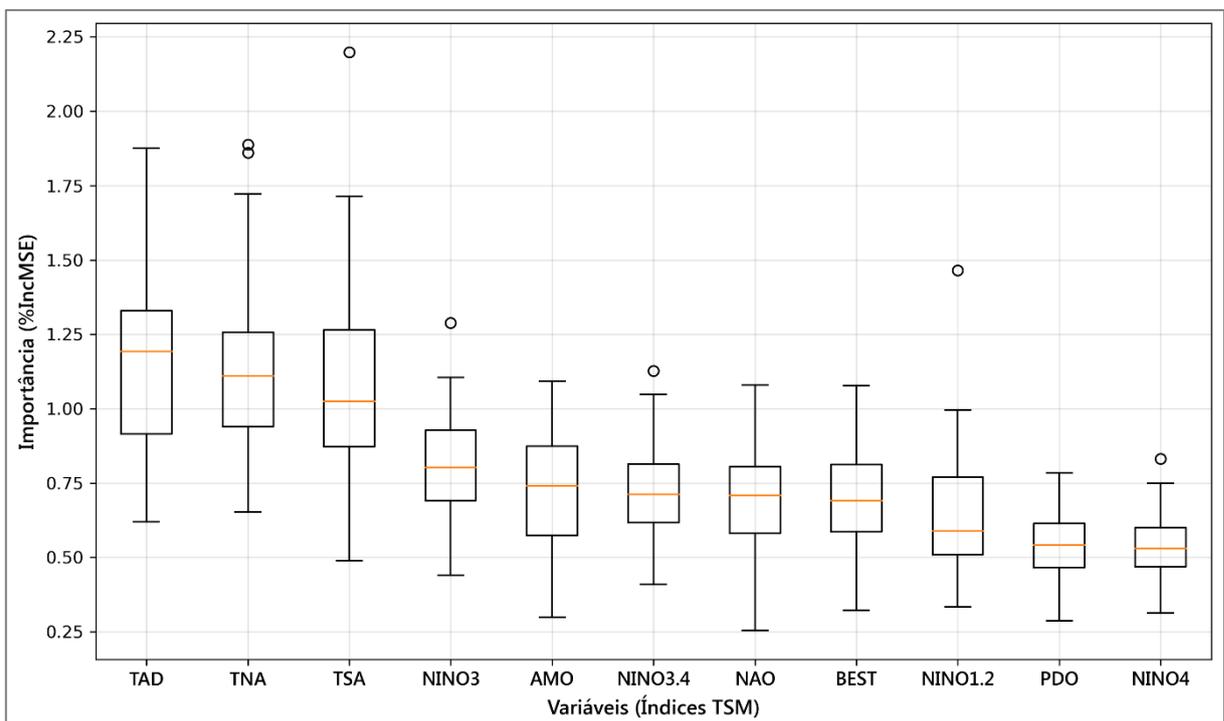
No modelo RLM (Figura 22), as métricas mostram tendências distintas entre treinamento e teste. O NSE de treinamento aumenta constantemente de M1 (0,2809) a M11 (0,4941), indicando melhoria na adequação do modelo com a adição de variáveis. No entanto, o NSE de validação começa positivo em M1 (0,2278), mas cai para valores negativos a partir de M3, piorando até M11 (-0,5052), um sinal de sobreajuste. A correlação de Pearson aumenta no treinamento até M11, mas declina no teste após M3. Estes resultados sugerem dificuldade de generalização no modelo de RLM, possivelmente devido à falta de uma regularização forte, levando a um aprendizado de ruídos ou padrões espúrios dos dados de treinamento que não se aplicam aos dados de validação.

4.3 Análise dos resultados para o período preditor MAM

4.3.1 Análise da importância dos índices TSM para o período preditor MAM

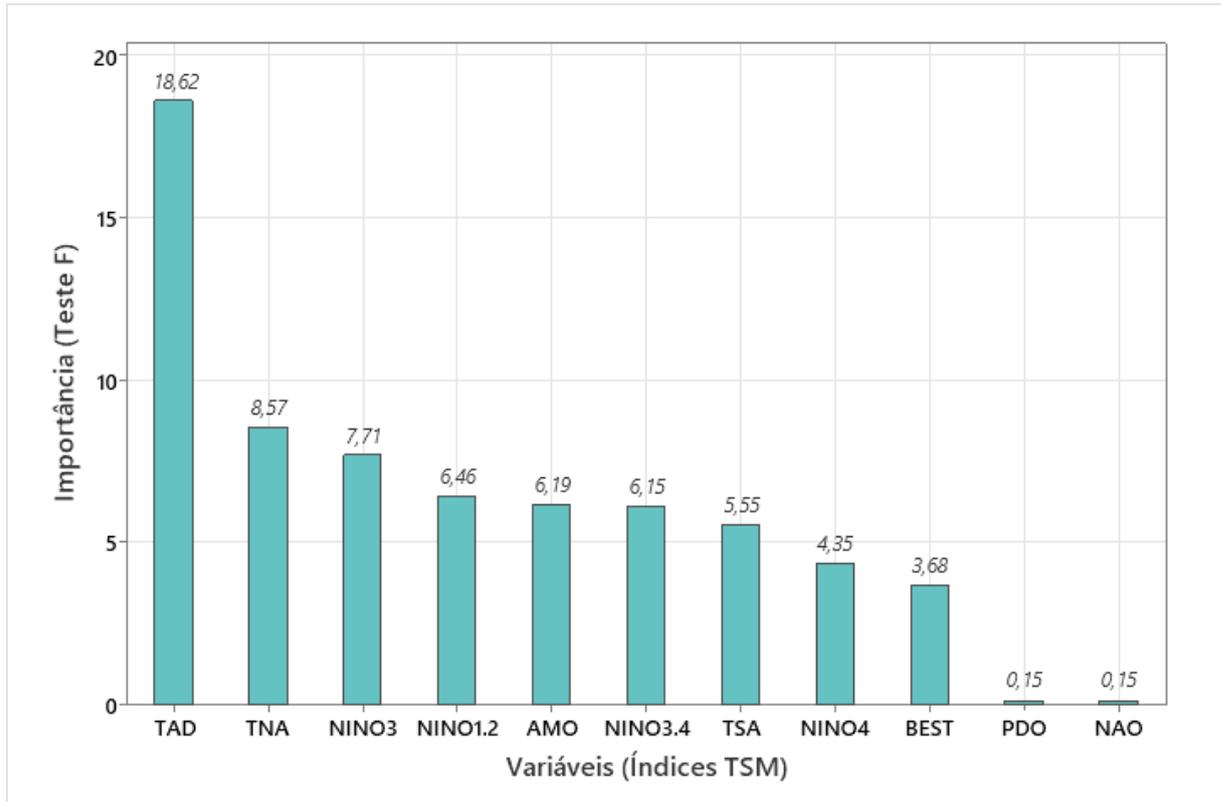
As 23, 24 e 25 apresentam a importância das variáveis para cada uma das técnicas utilizadas para esta avaliação, para o período preditor MAM.

Figura 23 - *Boxplot* da importância dos índices TSM no período preditor MAM (*Random Forest*)



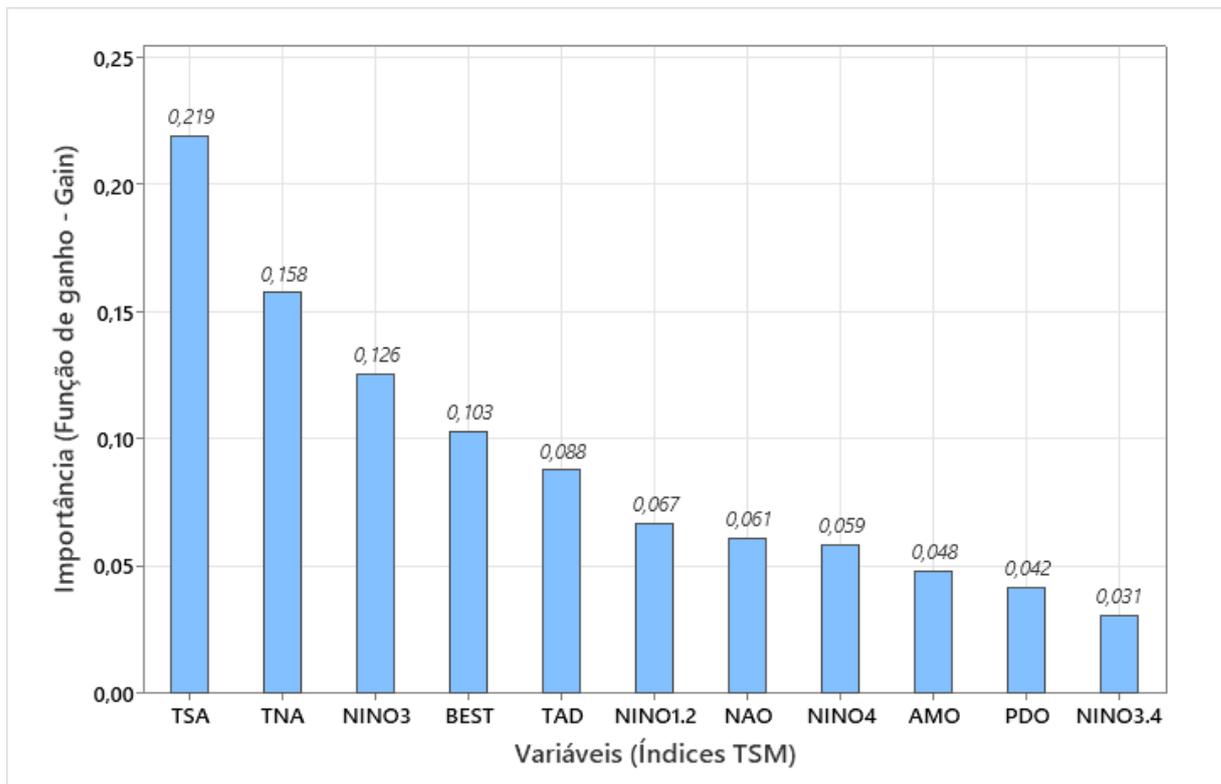
Fonte: Elaboração própria (2023).

Figura 24 - Importância dos índices TSM no período preditor MAM (*Select K Best*)



Fonte: Elaboração própria (2023).

Figura 25 - Importância dos índices TSM no período preditor MAM (*XGBoost*)



Fonte: Elaboração própria (2023).

A análise do trimestre MAM via *Random Forest*, apresentada na Figura 23, mostra um perfil mais equilibrado de importância entre as variáveis preditoras em comparação com períodos antecedentes. O índice TAD emerge como o mais influente, porém a proximidade das medidas de importância dos índices TNA e TSA sugere que estes também desempenham um papel significativo. A ocorrência de outliers em alguma das variáveis aponta para a existência de situações em que a importância desses índices diverge significativamente do comportamento geral, possivelmente devido a eventos climáticos excepcionais ou outras anomalias

Quando analisamos os resultados do *Select K Best* (Figura 24), o índice TAD também apresenta a maior importância, com um teste F de 18,62, seguido por TNA e NINO3. Este resultado sugere que o índice TAD mantém uma forte relação linear com a vazão durante o período de MAM, similar ao observado em outros trimestres, reforçando a sua relevância como uma boa variável TSM preditiva.

Na análise da Figura 25, os resultados do *XGBoost* para o trimestre MAM destacam o índice TSA como o mais influente, com um ganho de informação de 0,219, seguido por TNA e NINO3. Isso indica que, além da importância consistente do índice TAD obtida em outras abordagens, outros índices, como TSA e TNA, também desempenham papéis significativos na relação com a vazão anual para MAM. A presença do índice NINO3 como um preditor relevante pode apontar para efeitos específicos de fenômenos climáticos neste período preditor.

4.3.2 Ranqueamento unificado pelo método Copeland para o Preditor MAM

O Quadro 7, resultado da aplicação do método Copeland para unificar os rankings, apresenta um ranking unificado para este período preditivo, evidenciando as variáveis mais importantes baseado na integração dos resultados das abordagens anteriores.

Quadro 7 – *Ranking* Copeland para período MAM

Posição	Variável	Posição	Variável
1°	TNA	7°	BEST
2°	TAD	8°	NINO3.4
3°	NINO3	9°	NAO
4°	TSA	10°	NINO4
5°	AMO	11°	PDO
6°	NINO1.2		

Fonte: Elaboração própria (2023).

4.3.3 Modelos de previsão de vazão – MAM

As Tabelas 1 e 2, respectivamente, mostram a correlação e NSE entre vazões previstas e reais, nas fases de treinamento (T) e validação (V), para o período preditor MAM.

Tabela 1 - Avaliação do NSE Para o Preditor MAM

M's	MLP		LSTM		LASSO		RIDGE		RLM	
	T	V	T	V	T	V	T	V	T	V
M1	0,157	-0,008	0,081	-0,003	0,176	-0,107	0,156	-0,029	0,176	-0,107
M2	0,257	0,057	0,123	0,033	0,246	-0,041	0,174	0,034	0,252	-0,142
M3	0,284	-0,051	0,135	0,053	0,268	-0,020	0,196	0,053	0,278	-0,124
M4	0,248	0,022	0,122	0,024	0,269	-0,043	0,131	0,021	0,280	-0,084
M5	0,293	-0,022	0,118	0,011	0,265	-0,033	0,144	0,028	0,291	-0,211
M6	0,298	-0,126	0,129	0,045	0,271	-0,037	0,137	0,034	0,301	-0,250
M7	0,326	-0,084	0,105	0,022	0,271	-0,037	0,143	0,034	0,309	-0,249
M8	0,315	-0,054	0,122	0,026	0,272	-0,059	0,153	0,031	0,323	-0,398
M9	0,308	-0,271	0,098	0,001	0,262	-0,039	0,155	0,014	0,343	-0,451
M10	0,308	-0,209	0,092	0,016	0,262	-0,039	0,157	0,013	0,378	-0,904
M11	0,360	-0,271	0,101	-0,005	0,244	-0,016	0,157	0,012	0,383	-0,897

Fonte: Elaboração própria (2023).

Tabela 2 - Avaliação da Correlação de Pearson Para o Preditor MAM

M's	MLP		LSTM		LASSO		RIDGE		RLM	
	T	V	T	V	T	V	T	V	T	V
M1	0,426	0,128	0,419	0,124	0,412	0,120	0,412	0,120	0,412	0,120
M2	0,514	0,325	0,478	0,256	0,500	0,331	0,490	0,293	0,501	0,343
M3	0,553	0,274	0,504	0,307	0,521	0,347	0,512	0,315	0,526	0,367
M4	0,535	0,431	0,464	0,220	0,524	0,361	0,522	0,387	0,527	0,381
M5	0,548	0,364	0,440	0,207	0,522	0,359	0,505	0,375	0,538	0,319
M6	0,558	0,382	0,461	0,289	0,530	0,355	0,509	0,383	0,547	0,283
M7	0,576	0,338	0,430	0,246	0,530	0,355	0,508	0,363	0,554	0,271
M8	0,586	0,319	0,460	0,240	0,530	0,340	0,506	0,342	0,566	0,220
M9	0,572	0,298	0,446	0,236	0,535	0,318	0,512	0,311	0,581	0,207
M10	0,570	0,296	0,419	0,262	0,535	0,318	0,504	0,303	0,611	0,142
M11	0,611	0,284	0,417	0,227	0,523	0,319	0,502	0,299	0,615	0,141

Fonte: Elaboração própria (2023).

Na Tabelas 1, observa-se que os modelos MLP e LSTM mostraram um aumento no NSE de treino, com o MLP atingindo um pico em M11 (0,360) e o LSTM em M3 (0,135), indicando que ambos os modelos melhoraram sua adaptação aos dados de treino com a adição de variáveis. No entanto, o NSE de validação para ambos declinou substancialmente, sugerindo sobreajuste. Para o MLP, essa queda foi mais pronunciada, atingindo -0,271 em M11, enquanto para o LSTM a diminuição foi moderada, mas ainda resultou em um valor negativo em M11 (-0,004).

Lasso e o Ridge, modelos de regressão regularizada, também começaram com NSEs de treino fortes, mas seus NSEs de validação indicaram sobreajuste. O Lasso, especificamente, mostrou um NSE de validação consistentemente negativo, indicando que o modelo pode estar penalizando as variáveis de forma excessiva, o que afetou sua capacidade preditiva. O Ridge manteve valores de NSE de validação mais próximos de zero, sugerindo uma melhor generalização em comparação ao Lasso, mas ainda assim revelou um declínio na performance preditiva.

O RLM destacou-se com o maior NSE de treino, o que poderia indicar um ajuste muito bom aos dados de treino. Contudo, essa tendência não se refletiu nos dados de validação, onde o NSE foi decididamente negativo e diminuiu acentuadamente para -0,904 e -0,897 em M10 e M11, respectivamente. Esses valores negativos são indicativos de um modelo que está se ajustando excessivamente aos dados de treino e falhando em prever corretamente os dados de validação.

Analisando-se a Tabela 2, tem-se que o modelo MLP mostra uma tendência crescente na correlação de treino, partindo de 0,426 em M1 e alcançando 0,611 em M11, o que reflete um fortalecimento da relação linear à medida que mais variáveis são adicionadas. No entanto, para a validação, embora haja um aumento até M4 (0,431), há uma subsequente diminuição, sugerindo que as variáveis adicionais podem estar contribuindo para um sobreajuste.

Para o modelo LSTM, observamos um aumento na correlação de treino até M3 (0,504), seguido de uma diminuição em M4, o que pode indicar um ponto de inflexão na capacidade do modelo de capturar as relações lineares com a adição de variáveis. A correlação de validação atinge um pico em M2 (0,256) e depois diminui, o que é consistente com a tendência observada no NSE, onde o modelo começa a perder a capacidade de generalização após M2.

O Lasso e o Ridge apresentam um padrão de correlação de treino semelhante, com um crescimento até M3 (Lasso: 0,521, Ridge: 0,512) e, então, uma estabilidade ou leve

flutuação. A correlação de validação para ambos os modelos mostra um pico em M3, após o qual há uma diminuição. Este comportamento paralelo sugere que a adição de variáveis até M3 melhora a relação linear, mas adicionais variáveis não contribuem de forma significativa, possivelmente devido à penalização das variáveis menos informativas.

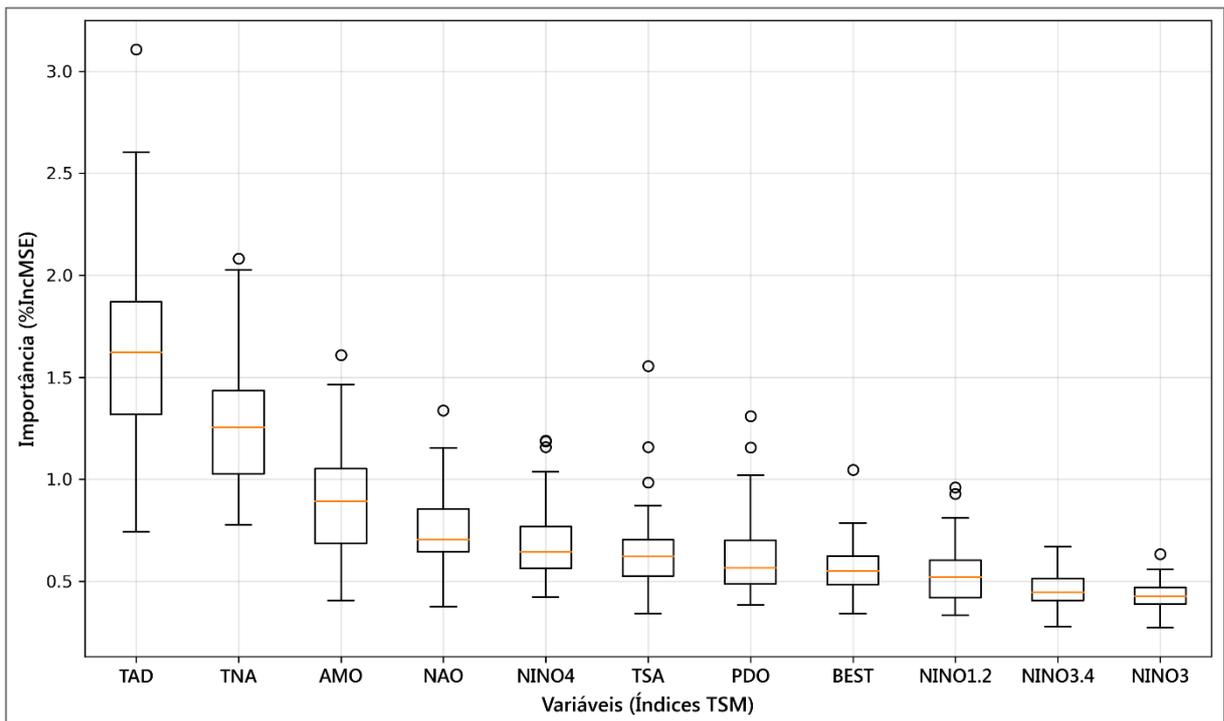
O RLM exibe um padrão distintivo com a maior correlação de treino, alcançando 0,615 em M11, indicando um ajuste muito forte aos dados de treino. No entanto, a correlação de validação não segue essa tendência, começando em 0,120 em M1 e caindo para 0,141 em M11. Isso sugere que enquanto o modelo se ajusta bem aos dados de treino, ele não é capaz de manter essa performance em dados não vistos, evidenciando uma vez mais o sobreajuste, como foi fortemente indicado pelos resultados do NSE.

4.4 Análise dos resultados para o período preditor DJF

4.4.1 Análise da importância dos índices TSM para o período preditor DJF

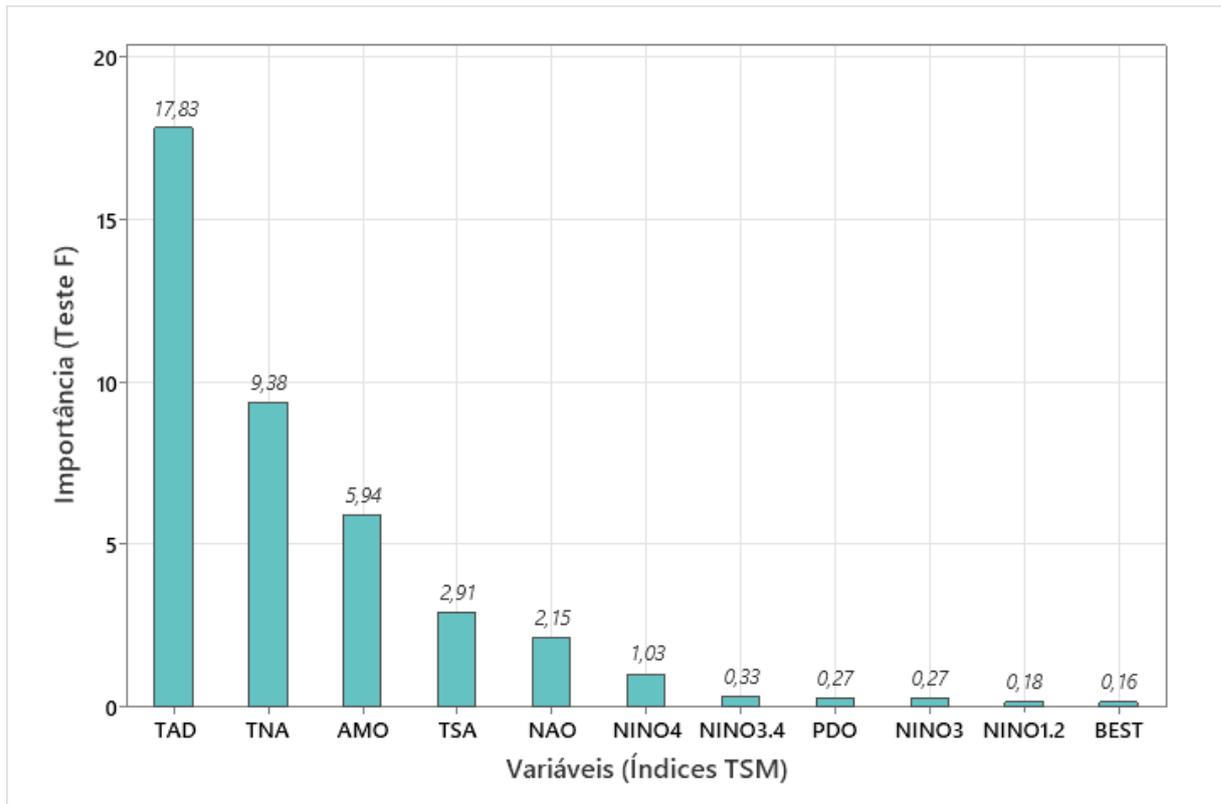
As 26, 27 e 28 apresentam a importância das variáveis para cada uma das técnicas utilizadas para esta avaliação, para o período preditor DJF.

Figura 26 - *Boxplot* da importância dos índices TSM no período preditor DJF (*Random Forest*)



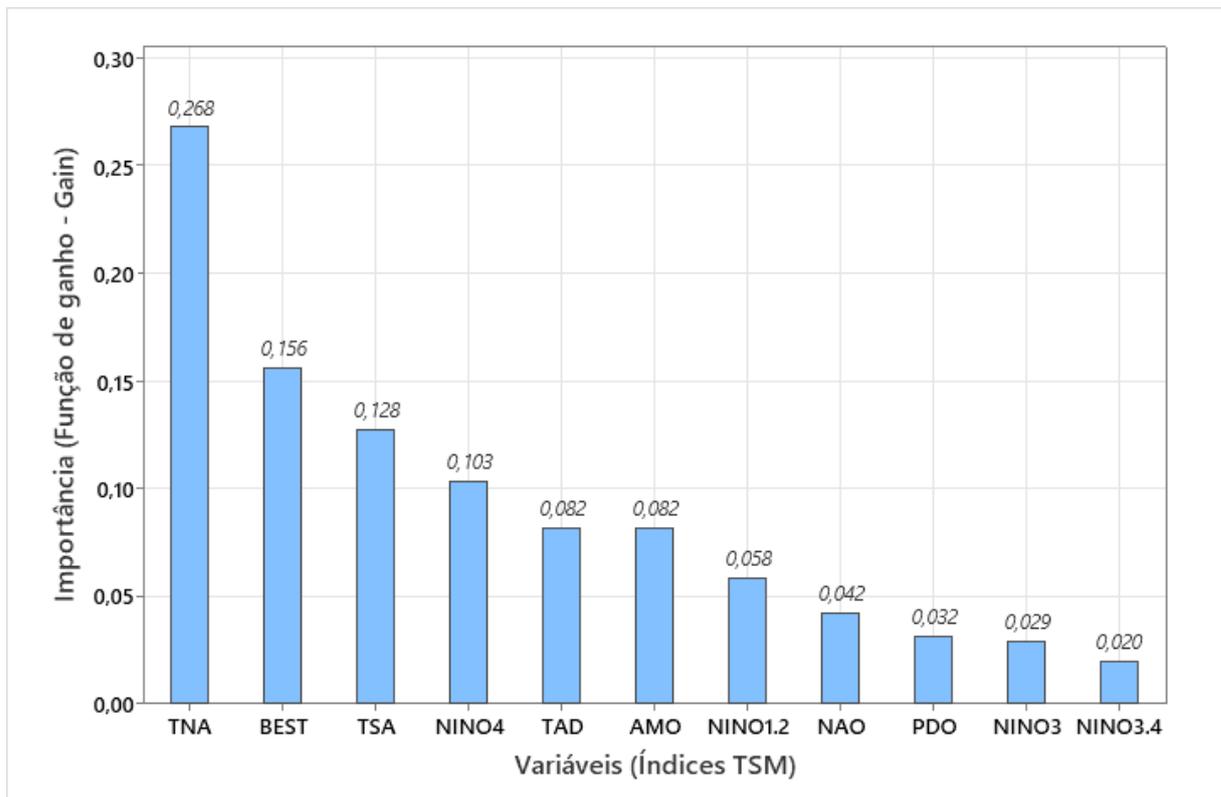
Fonte: Elaboração própria (2023).

Figura 27 - Importância dos índices TSM no período preditor DJF (*Select K Best*)



Fonte: Elaboração própria (2023).

Figura 28 - Importância dos índices TSM no período preditor DJF (*XGBoost*)



Fonte: Elaboração própria (2023).

O *Random Forest* aplicado ao período preditor de DJF indica que o índice TAD detém a maior importância, conforme mostrado pelo gráfico de *boxplot* (Figura 26). Este índice exibe uma variabilidade significativa, sugerindo que ele pode ter um papel fundamental e variável na previsão de vazão no trimestre DJF. Outros índices como TNA e AMO também mostram importância, mas são menos variáveis, indicando uma influência mais consistente no período. Além disso, observa-se que o índice TSA apresenta importância relativamente menor comparado a sua colocação e medida de importância em outros períodos preditores.

A análise com o *Select K Best* (Figura 27) para o mesmo trimestre realça novamente o índice TAD como o mais significativo com um teste F de 17,83, seguido de TNA e AMO, com 9,38 e 5,94 respectivamente. Esta consistência do índice TAD como uma variável importante em diferentes métodos reforça a sua posição como variável preditora chave para a vazão durante o trimestre DJF.

Os resultados do *XGBoost* (Figura 28), por sua vez, divergem ligeiramente, apresentando o índice TNA como o mais influente com um ganho de informação de 0,268. A importância do índice TAD é menos destacada neste método em comparação com o *Select K Best*, mas ainda é relevante. O índice BEST também se mostra como um preditor significativo com um ganho de 0,156, sugerindo que variáveis adicionais devem ser consideradas ao prever a vazão neste trimestre.

4.4.2 Ranqueamento unificado pelo método Copeland para o Preditor DJF

Semelhante aos períodos anteriores, o Quadro 8, derivada da implementação do método Copeland para a síntese de classificações distintas, revela uma classificação unificada para o período preditor JJA, ressaltando as variáveis-chave com base na combinação dos dados obtidos das técnicas antecedentes.

Quadro 8 – *Ranking* Copeland para período DJF

Posição	Variável	Posição	Variável
1°	TNA	7°	BEST
2°	TAD	8°	NINO3.4
3°	NINO3	9°	NAO
4°	TSA	10°	NINO4
5°	AMO	11°	PDO
6°	NINO1.2		

Fonte: Elaboração própria (2023).

4.4.3 Modelos de previsão de vazão – DJF

As Tabelas 3 e 4, respectivamente, mostram a correlação e NSE entre vazões previstas e reais, nas fases de treinamento (T) e validação (V), para o período preditor DJF.

Tabela 3 - Avaliação do NSE Para o Preditor DJF

M's	MLP		LSTM		LASSO		RIDGE		RLM	
	T	V	T	V	T	V	T	V	T	V
M1	0,164	0,007	0,096	0,006	0,167	-0,070	0,162	-0,032	0,167	-0,070
M2	0,227	0,085	0,130	0,060	0,214	-0,014	0,174	0,061	0,215	-0,021
M3	0,249	0,106	0,127	0,044	0,211	-0,019	0,162	0,042	0,217	-0,054
M4	0,290	-0,003	0,118	0,019	0,205	-0,014	0,147	0,027	0,246	-0,127
M5	0,293	0,055	0,129	0,019	0,219	-0,020	0,134	0,019	0,287	-0,219
M6	0,296	0,000	0,087	0,009	0,236	-0,023	0,136	0,014	0,312	-0,321
M7	0,314	-0,009	0,092	-0,004	0,236	-0,023	0,139	0,011	0,316	-0,344
M8	0,323	-0,002	0,096	-0,015	0,236	-0,023	0,140	0,007	0,318	-0,360
M9	0,387	-0,012	0,103	-0,017	0,236	-0,023	0,141	0,004	0,325	-0,579
M10	0,434	-0,055	0,118	-0,020	0,236	-0,023	0,146	-0,002	0,332	-0,550
M11	0,336	-0,059	0,094	-0,049	0,236	-0,023	0,149	-0,007	0,343	-0,537

Fonte: Elaboração própria (2023).

Tabela 4 - Avaliação da Correlação de Pearson Para o Preditor DJF

M's	MLP		LSTM		LASSO		RIDGE		RLM	
	T	V	T	V	T	V	T	V	T	V
M1	0,436	0,155	0,416	0,146	0,400	0,130	0,400	0,130	0,400	0,130
M2	0,482	0,320	0,481	0,326	0,462	0,330	0,458	0,314	0,462	0,333
M3	0,513	0,357	0,486	0,322	0,462	0,325	0,438	0,287	0,464	0,324
M4	0,541	0,326	0,464	0,242	0,461	0,315	0,443	0,311	0,492	0,302
M5	0,557	0,363	0,505	0,251	0,486	0,302	0,471	0,292	0,530	0,299
M6	0,551	0,291	0,390	0,216	0,505	0,308	0,493	0,278	0,551	0,297
M7	0,567	0,310	0,419	0,187	0,505	0,308	0,497	0,270	0,555	0,287
M8	0,586	0,317	0,422	0,156	0,505	0,308	0,496	0,263	0,556	0,281
M9	0,626	0,352	0,447	0,166	0,505	0,308	0,498	0,256	0,564	0,207
M10	0,664	0,335	0,461	0,160	0,505	0,308	0,501	0,242	0,569	0,188
M11	0,599	0,306	0,406	0,172	0,505	0,308	0,502	0,232	0,578	0,194

Fonte: Elaboração própria (2023).

A partir da análise da Tabela 3, tem-se que para o MLP, nota-se um crescimento no NSE de treino, atingindo um pico em M10 (0,434), que sugere uma melhora na habilidade do modelo de capturar a variância nos dados de treino com a inclusão de mais variáveis. No entanto, o NSE de validação apresenta um pico em M2 (0,085) e depois cai significativamente, indicando problemas de sobreajuste conforme o modelo se torna mais complexo.

O LSTM tem um perfil similar no treinamento com um pico em M2 (0,130), mas a validação mostra uma tendência menos clara. O NSE positivo em M2 (0,060) sugere um bom ajuste inicial, mas a métrica torna-se negativa a partir de M11 (-0,049), o que pode ser um sinal de sobreajuste ou de inadequação do modelo para capturar a estrutura temporal nos dados de validação.

Os modelos de Regressão Lasso e Ridge apresentam um comportamento estável no treinamento, mas com NSEs de validação consistentemente negativos, começando em -0,070 e -0,032, respectivamente. Isso implica que, embora os modelos estejam se ajustando aos dados de treino, eles não estão generalizando bem, possivelmente devido a um encolhimento excessivo das estimativas de coeficientes.

O RLM mostra um aumento no NSE de treino até M11 (0,343), sugerindo um ajuste crescente aos dados de treino. No entanto, semelhante aos modelos de regressão regularizada, o NSE de validação para o RLM é negativo em todos os modelos, indicando que o modelo está sobreajustando aos dados de treino e perdendo a capacidade de generalização.

Na análise entre os resultados de Correlação de Pearson (Tabela 4) para os modelos MLP, LSTM, Lasso, Ridge e RLM, percebemos um padrão de comportamento que espelha as tendências observadas com a métrica NSE. O modelo MLP demonstra um aumento na correlação de treino com um pico em M10, sugerindo um bom ajuste inicial aos dados de treino. No entanto, a correlação de validação diminui a partir de M2, indicando um possível sobreajuste.

O LSTM segue um padrão semelhante ao MLP em treino, com uma diminuição subsequente na correlação de validação, reforçando a preocupação com a capacidade de generalização do modelo. Os modelos Lasso e Ridge mostram uma maior estabilidade na correlação tanto no treino quanto na validação, implicando uma melhor generalização em comparação ao MLP e LSTM.

O RLM, apesar de ter um aumento constante na correlação de treino, enfrenta desafios significativos de sobreajuste, como evidenciado pela diminuição da correlação de validação. Isso reafirma a necessidade de atenção ao equilibrar a complexidade do modelo e

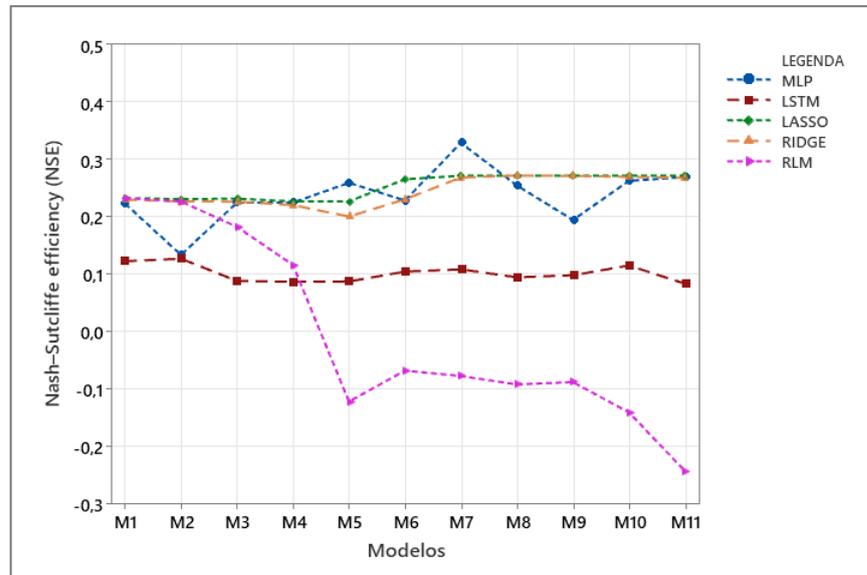
sua habilidade de generalizar, para evitar a adaptação excessiva aos dados de treino que não se traduz em um bom desempenho preditivo em dados novos.

4.5 Comparação e Síntese dos Resultados

Nesta seção, destaca-se a análise de desempenho para os modelos dos períodos preditores SON e JJA, selecionados por apresentarem os resultados mais promissores nas etapas de validação. Além disso, esta análise visa elencar os modelos variantes que obtiveram os melhores resultados, em termo de ganhos de desempenho para cada abordagem de previsão, enfatizando a relevância destes no contexto de cada preditor. Dessa forma, optou-se por não incluir os preditores MAM e DJF nesta análise, considerando os indícios de sobreajuste e desempenhos inconsistentes em seus modelos, como evidenciado pelos valores negativos de NSE e correlação fraca nas Tabelas 1, 2, 3 e 4.

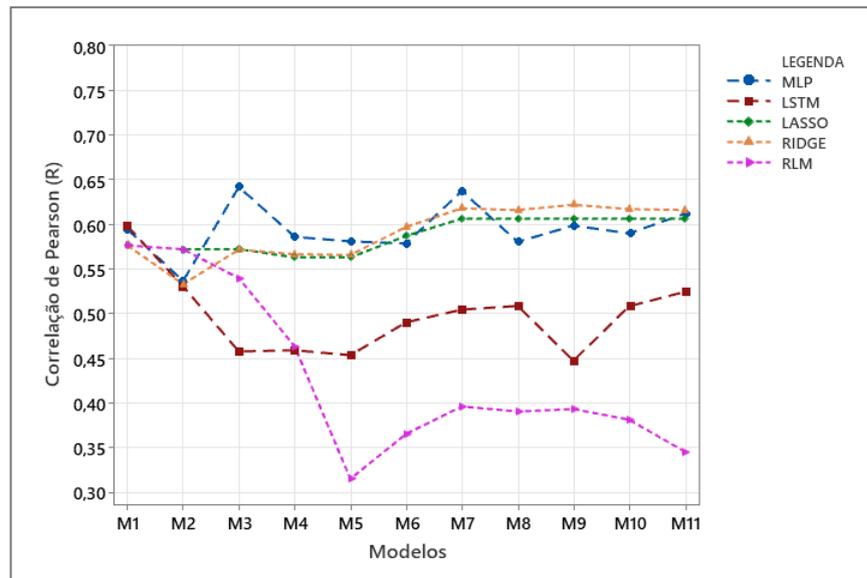
As Figuras 29 e 30 apresentam a comparação direta entre cada um dos modelos do preditor SON para as métricas de avaliação na etapa de validação. Já as Tabelas 5 e 6, destacam a variação percentual dos modelos em comparação ao desempenho do modelo variante M1 para cada uma das modelagens.

Figura 29 – Comparação entre os Modelos de SON (NSE - validação)



Fonte: Elaboração própria (2023).

Figura 30 – Comparação entre os Modelos de SON (R - Validação)



Fonte: Elaboração própria (2023).

Tabela 5 – Variação Percentual de Modelos de SON vs. M1 (NSE - Validação)

Modelos	MLP	LSTM	LASSO	RIDGE	RLM
M1	0%	0%	0%	0%	0%
M2	-40%	3%	-1%	-1%	-3%
M3	1%	-28%	0%	-2%	-22%
M4	1%	-29%	-3%	-4%	-50%
M5	16%	-29%	-3%	-13%	-152%
M6	2%	-15%	14%	0%	-129%
M7	47%	-12%	17%	17%	-133%
M8	14%	-23%	17%	18%	-139%
M9	-13%	-20%	17%	18%	-138%
M10	18%	-6%	17%	17%	-160%
M11	21%	-32%	17%	17%	-204%

Fonte: Elaboração própria (2023).

Tabela 6 – Variação Percentual de Modelos de SON vs. M1 (R - Validação)

Modelos	MLP	LSTM	LASSO	RIDGE	RLM
M1	0%	0%	0%	0%	0%
M2	-10%	-11%	-1%	-8%	-1%
M3	8%	-24%	-1%	-1%	-6%
M4	-1%	-23%	-2%	-2%	-20%
M5	-2%	-24%	-2%	-2%	-45%
M6	-3%	-18%	2%	4%	-37%
M7	7%	-16%	5%	7%	-31%
M8	-2%	-15%	5%	7%	-32%
M9	1%	-25%	5%	8%	-32%
M10	-1%	-15%	5%	7%	-34%
M11	3%	-12%	5%	7%	-40%

Fonte: Elaboração própria (2023).

A análise detalhada das Figuras 29 e 30 permite a identificação dos modelos que apresentam os melhores desempenho, destacando-se as variações nas métricas ao incrementar as variáveis preditoras. Notavelmente, o modelo MLP emerge como superior em termos de métricas NSE e correlação. Embora o modelo MLP demonstre oscilações, observa-se uma tendência de crescimento consistente, particularmente nos modelos M3, M5 e M7.

Em comparação, as regressões Lasso e Ridge também exibem um incremento no desempenho, caracterizado por uma evolução mais uniforme dos modelos M1 a M5, culminando em um aumento notável em M7. Notoriamente, o modelo Ridge exibe uma ligeira preeminência sobre o modelo Lasso na métrica NSE, conforme ilustrado na Figura X.

Adicionalmente, a análise da Tabela 5 revela que o modelo MLP-M7 distingue-se com um incremento percentual de 47% na eficiência de NSE, comparado ao modelo base M1. Esta elevação sugere uma capacidade aprimorada de previsão do modelo. Por outro lado, o modelo LSTM apresenta uma tendência de declínio na métrica NSE, com uma redução notável de até 32% em M11, o que pode indicar uma incompatibilidade das modificações aplicadas com a estrutura do modelo.

Da mesma forma, os modelos Lasso e Ridge exibem inicialmente um comportamento mais estável, com pequenas variações negativas, entretanto, o cenário muda a partir de M6 com variações percentuais positivas atingindo um ganho percentual de 17% em M7, indicando uma consistência na eficiência das previsões. Entretanto, o modelo RLM apresenta um declínio acentuado, com uma variação percentual negativa de até -204% em M11, destacando um desempenho ineficiente.

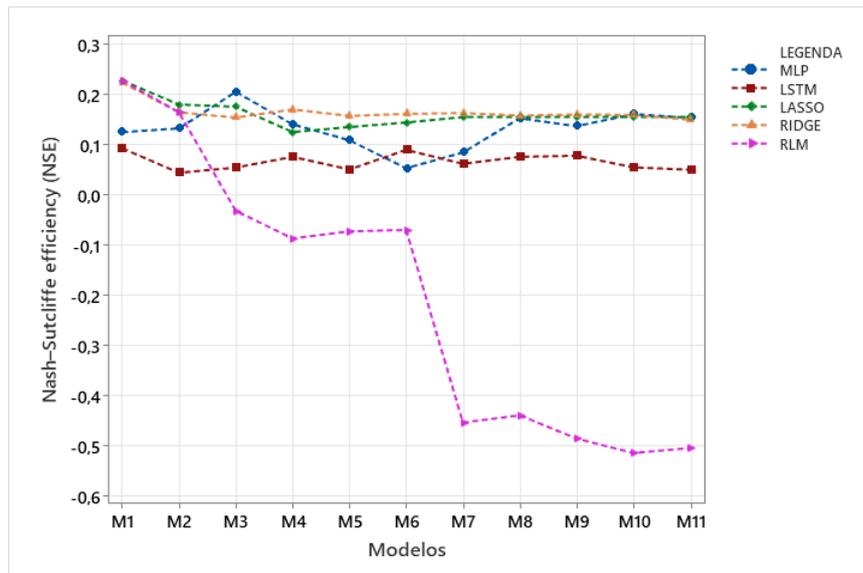
A análise da variação percentual para a correlação (Tabela 6) reitera a performance estável do MLP, com melhorias observadas em variantes como M3 e M7, com ganhos de 8 e 7%, respectivamente. O modelo Ridge, por sua vez, se destaca por uma melhoria gradual em suas variantes mais recentes, enquanto o Lasso mantém uma consistência próxima ao modelo base. Contrastando, o LSTM, apesar de um início promissor, revela uma diminuição generalizada em suas variantes, assim como modelo RLM que sofre uma degradação notável na correlação de Pearson, à medida que é adicionado mais variáveis.

Em síntese, o modelo MLP, especialmente na variante M7, consolida-se como o mais eficiente e consistente em ambas as métricas avaliadas, se mostrando como o melhor modelo de previsão associado ao período preditor SON. Por outro lado, os modelos Lasso e Ridge também se mostram como opções confiáveis, com variações positivas estáveis em ambas as métricas. No entanto, os resultados relativos ao LSTM e ao RLM sugerem que para esta

quantidade de dados e configurações, os modelos não conseguem capturar a dinâmica das variáveis e realizar previsões eficientes.

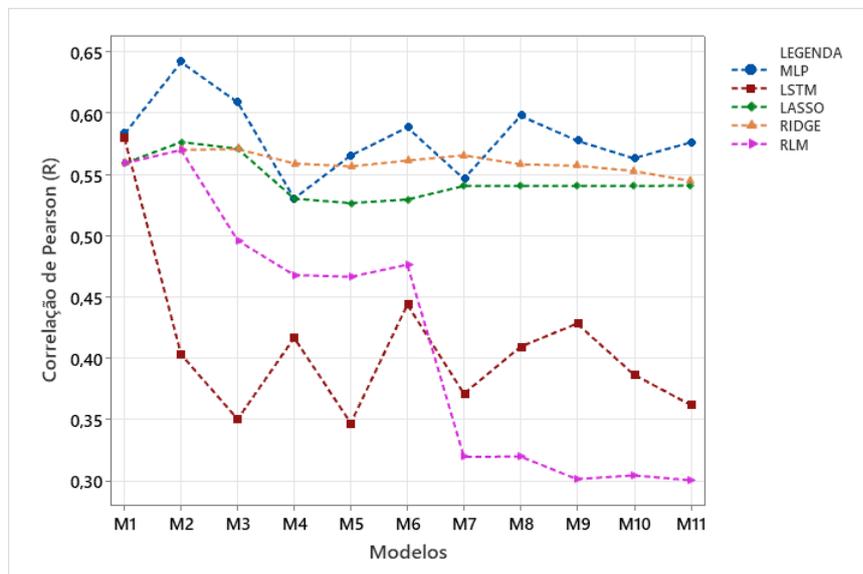
De forma análoga à abordagem utilizada para analisar os modelos associados ao período preditor SON, as Figuras 31 e 32, juntamente das Tabelas 7 e 8, oferecem uma visão comparativa detalhada para o preditor JJA. Essa abordagem possibilita uma compreensão das características e do desempenho de cada modelo, destacando suas particularidades no contexto do preditor JJA.

Figura 31 – Comparação entre os modelos de JJA (NSE - validação)



Fonte: Elaboração própria (2023).

Figura 32 – Comparação entre os modelos de JJA (R - validação)



Fonte: Elaboração própria (2023).

Tabela 7 – Variação Percentual dos Modelos de JJA vs. M1 (NSE - Validação)

Modelos	MLP	LSTM	LASSO	RIDGE	RLM
M1	0%	0%	0%	0%	0%
M2	6%	-53%	-21%	-27%	-28%
M3	64%	-42%	-23%	-31%	-114%
M4	13%	-19%	-45%	-24%	-138%
M5	-13%	-46%	-41%	-30%	-132%
M6	-58%	-4%	-37%	-28%	-131%
M7	-32%	-33%	-32%	-27%	-299%
M8	22%	-18%	-32%	-29%	-293%
M9	10%	-16%	-32%	-29%	-313%
M10	29%	-41%	-32%	-29%	-326%
M11	23%	-46%	-32%	-33%	-322%

Fonte: Elaboração própria (2023).

Tabela 8 – Variação Percentual dos Modelos de JJA vs. M1 (R - Validação)

Modelos	MLP	LSTM	LASSO	RIDGE	RLM
M1	0%	0%	0%	0%	0%
M2	-10%	-11%	-1%	-8%	-1%
M3	8%	-24%	-1%	-1%	-6%
M4	-1%	-23%	-2%	-2%	-20%
M5	-2%	-24%	-2%	-2%	-45%
M6	-3%	-18%	2%	4%	-37%
M7	7%	-16%	5%	7%	-31%
M8	-2%	-15%	5%	7%	-32%
M9	1%	-25%	5%	8%	-32%
M10	-1%	-15%	5%	7%	-34%
M11	3%	-12%	5%	7%	-40%

Fonte: Elaboração própria (2023).

A análise da Figura 31 evidencia uma performance notável do modelo MLP, particularmente na variante M3. Esta observação é corroborada pela Tabela 7, que registra um incremento significativo de 64% na métrica NSE para o modelo MLP-M3. Apesar deste pico de eficiência, variações nos desempenhos de outros modelos foram notadas, manifestando-se em flutuações.

Diferentemente, o modelo LSTM, exibe um perfil de desempenho mais uniforme, ainda que com oscilações. Contudo, a Tabela 7 revela uma tendência geralmente negativa na variação percentual de NSE, indicando uma diminuição na precisão das previsões. Por outro lado, as regressões Lasso e Ridge mostram estabilidade em suas variações percentuais, apesar de uma tendência negativa contínua após a variante M7. Em comparação, o modelo de Regressão Linear Múltipla (RLM) apresenta um desempenho substancialmente inferior, com variações percentuais negativas pronunciadas, sugerindo uma inadequação para o conjunto de dados analisado.

Avaliando a correlação de Pearson na Figura 32 e Tabela 8, o modelo MLP demonstra uma variação menos acentuada, mantendo-se relativamente estável com melhorias marginais em algumas variantes. Em contrapartida, o LSTM registra uma perda substancial na manutenção da correlação linear, evidenciada por reduções significativas em diversas variantes. Lasso e Ridge exibem um comportamento mais consistente, com o modelo Ridge mantendo uma correlação próxima ao valor base em todas as suas variantes. Novamente, o RLM destaca-se negativamente, evidenciando uma degradação progressiva em sua capacidade de manter uma correlação linear robusta.

Em resumo, o modelo MLP, especialmente na variante M3, emerge como o mais promissor em termos de eficiência preditiva para o período preditor JJA. Por outro lado, o LSTM mostra uma capacidade reduzida de adaptação e aprimoramento. Os modelos de regressão Lasso e Ridge, embora estáveis, exibem uma eficiência ligeiramente decrescente. O RLM, por sua vez, revela-se o menos apropriado, com um desempenho declinante tanto em NSE quanto na correlação de Pearson, similarmente ao ocorrido no período preditor SON.

5 CONCLUSÃO

O objetivo principal deste trabalho foi classificar os índices de Temperatura de Superfície do Mar (TSM) segundo a sua influência nas vazões anuais afluentes ao reservatório Fronteiras. Para tal, foi realizada uma análise detalhada da importância dessas variáveis, empregando três métodos de aprendizado de máquina: *Random Forest*, *Select K Best* e *XGBoost*. Esses métodos contribuíram para a compilação de um ranking unificado, estabelecido através do método de Copeland. Subsequentemente, com base neste ranking, foram desenvolvidos cinco modelos preditivos de vazão, cada um apresentando 11 variações distintas, por meio do uso de Redes Neurais Artificiais (RNAs) e modelos de regressão Linear Múltipla. Este trabalho, portanto, possibilitou uma análise em relação à compreensão da influência dos índices TSM nas vazões anuais e, além disso, permitiu explorar metodologias de previsão de vazão em reservatórios.

Em relação à importância dos índices de Temperatura de Superfície do Mar (TSM), o Dipolo do Atlântico Tropical (TAD) e as anomalias de TSM associadas a ele, especificamente o Atlântico Tropical Norte e o Atlântico Tropical Sul, foram identificados como os mais influentes. Esta classificação se manteve consistente nos três métodos de aprendizado de máquina utilizados para a seleção de variáveis em todos os trimestres preditores. Consequentemente, estando entre os três primeiros no ranking de Copeland, com a exceção do índice TSA, que ocupou a quarta posição nos períodos preditores MAM e DJF.

Adicionalmente às variáveis anteriormente mencionadas, outras anomalias, como a Oscilação Multidecadal do Atlântico (AMO) e a Oscilação do Atlântico Norte (NAO), também demonstraram sua relevância. Estas foram classificadas entre as sete variáveis mais influentes no ranking de Copeland. É importante destacar, no entanto, que o NAO apresentou uma posição relativamente inferior nos trimestres MAM e DJF, divergindo do padrão observado nos períodos preditores SON e JJA. Além destas, o índice NINO1.2 também se destacou, posicionando-se consistentemente entre as sete variáveis mais relevantes no ranking de Copeland para todos os períodos preditores avaliados.

Quanto aos modelos de previsão de vazão, tem-se que SON e JJA se mostraram como sendo os períodos preditores em que os modelos associados apresentaram os melhores desempenhos. Isso ocorreu dentro deste contexto em que a quantidade de dados (amostras) e a estrutura padrão dos modelos foram mantidas fixas, sem alterações em função do trimestre preditor.

Diante dos resultados obtidos, destaca-se o modelo SON-MLP-M7, que alcançou um NSE de 0,329 e uma correlação de 0,638 na etapa de validação. Já o modelo SON-Lasso-M7 registrou um NSE de 0,272 e uma correlação de 0,607, enquanto a Regressão SON-Ridge-M7 obteve um NSE de 0,269 e correlação de 0,618 na validação.

Para o período JJA, sobressaem os modelos JJA-MLP-M3, com NSE de 0,205 e correlação de 0,609, e as Regressões Lasso, Ridge e RLM associadas ao M1, que apresentaram NSE aproximadamente 0,23 e correlação em torno de 0,55 na etapa de validação. É importante destacar que, no contexto desta pesquisa, embora os modelos LSTM e RLM tenham apresentado resultados promissores nas métricas de treinamento com a adição de mais variáveis, eles não foram eficazes em capturar de maneira adequada o comportamento dos dados na etapa de validação. Isso levou a um desempenho insatisfatório, indicando limitações desses modelos dentro das condições e particularidades deste estudo.

Portanto, ao analisar os resultados de maneira abrangente, identifica-se que os modelos mais eficientes correspondem ao período preditor SON, com destaque para o modelo MLP-M7. Inicialmente, este modelo, associado às variáveis TAD, TNA e TSA, já apresentou desempenhos consideráveis. No entanto, observou-se um avanço ainda mais expressivo em termos de variação percentual no desempenho ao incluir um conjunto mais amplo, com sete variáveis, que inclui TAD, TNA, TSA, BEST, AMO e NINO1.2.

Levando em conta o tamanho da série histórica, e as características desta pesquisa, os resultados alcançados em termos de correlação podem ser considerados satisfatórios. No entanto, os melhores modelos ainda apresentaram valores de NSE relativamente baixos. Portanto, com o objetivo de aprimorar os estudos sobre previsão de vazão, sugere-se para futuras pesquisas:

- a) Utilizar outros métodos de seleção e classificação de variáveis.
- b) Utilizar a própria variável de vazão como dado de entrada nos modelos;
- c) Utilizar a variável de precipitação como dado de entrada nos modelos;
- d) Avaliar todas as combinações possíveis de variáveis TSMs;
- e) Realizar a calibração de parâmetros dos modelos para cada período preditor.

REFERÊNCIAS

ABADI, M. *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. **arXiv preprint arXiv:1603.04467**, 2016.

AGÊNCIA NACIONAL DE ÁGUAS (Brasília, DF). Centro de Gestão e Estudos Estratégicos. **A Questão da Água no Nordeste**. [S. l.: s. n.], 2012.

AMORIM, G. B. C. de. **Random-subset voting**. 2020. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Pernambuco, Recife, 2020.

ANDREOLI, R. V.; KAYANO, M. T. A importância relativa do Atlântico Tropical Sul e Pacífico Leste na variabilidade de precipitação do Nordeste do Brasil. **Revista Brasileira de Meteorologia**, v. 22, p. 63-74, 2007.

ANDREOLI, R. V.; KAYANO, M. T. Multi-scale variability of the sea surface temperature in the Tropical Atlantic. **Journal of Geophysical Research C: Oceans**, v. 109, n. 5, p. 1–12, 2004.

ANJOS, S. L. **Ondaletas aplicadas à análise de variabilidade de baixa frequência em séries de afluição aos reservatórios hidrelétricos brasileiros**. Dissertação (Mestre) em Tecnologia Ambiental e Recursos Hídricos, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, 2015.

ARAGÃO, J. O. R. de. O Impacto do ENSO e do Dipolo do Atlântico no Nordeste do Brasil. **Bulletin de l'Institut Français d'Études Andines**, v. 27, n. 3, p. 839-844, 1998.

ARAÚJO JÚNIOR, L. M. A.; SOUZA FILHO; F. A. S.; BARRETO, G A. **Previsão de vazão: Multi modelos**. In: ADAPTA: Gestão adaptativa do risco climático de seca/ organizado por Francisco de Assis de Souza Filho *et al.* Fortaleza: Expressão Gráfica e Editora, 2019.

ARAÚJO, C. B. C de. *et al.* Previsão Sazonal de Vazões para a Bacia do Orós (Ceará, Brasil) Utilizando Redes Neurais e a Técnica De Reamostragem dos K-vizinhos. **Revista Brasileira de Meteorologia**, Fortaleza, v. 35, p. 197-207, 2020.

ARAÚJO, C. B. C. **Aplicação das redes neurais artificiais do tipo perceptron na estimativa de recalques em estacas**. 2015. 203 f. Dissertação (Mestrado em Engenharia Civil: Geotecnia) -Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza, 2015.

ARAÚJO, C. B. C. de; DANTAS NETO, S. A.; SOUZA FILHO, F. de A. Previsão de vazões para o açude Orós/CE a partir de dados hidrometeorológicos utilizando perceptrons. **Revista Brasileira de Meteorologia**, Fortaleza, v. 30, n. 1, p. 37-46, 2015.

ARAÚJO, C. B. C. *et al.* **Previsão de vazões: Utilização de RNA**. In: ADAPTA: Gestão adaptativa do risco climático de seca/ organizado por Francisco de Assis de Souza Filho *et al.* Fortaleza: Expressão Gráfica e Editora, 2019.

ARAÚJO, M. S. da. *et al.* Metodologia baseada em redes neurais artificiais para previsão da geração de plantas fotovoltaicas. In: **Congresso Brasileiro de Energia Solar-CBENS**. 2022. p. 1-10.

BARAN, C. T.; SANTOS, M. S. Comparação entre Modelos Estatísticos, Arima e Técnicas de Machine Learning para Previsão de Vazões em Porto Velho, Bacia do Rio Madeira. **XXIII SBRH - Simpósio Brasileiro de Recursos Hídricos**, Foz do Iguaçu - PR, 2019.

BARBA-ROMERO, S.; POMEROL, J-C. **Decisiones multicriterio: fundamentos teóricos y utilización práctica**. Universidad de Alcalá, Servicio de Publicaciones, 1997.

BARKER, J. *et al.* Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. **Medical image analysis**, v. 30, p. 60-71, 2016.

BATISTA, A. L. F. **Modelos de Séries Temporais e Redes Neurais Artificiais na Previsão de Vazão**. Dissertação (Mestrado em Modelagem de Sistemas Biológicos) - Universidade Federal de Lavras- MG. Lavras- MG, 2009.

BIRIKUNDAVYI, S. *et al.* Performance of neural networks in daily streamflow forecasting. **Journal of Hydrologic Engineering**, v. 7, n. 5, p. 392-398, 2002.

BISHOP, C. M.; NASRABADI, Nasser M. **Pattern recognition and machine learning**. New York: springer, 2006.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais – Teoria e Aplicações**. 2ª edição. Editora LTC, 2007.

BRAVO, J. M. *et al.* Avaliação dos benefícios de previsões de vazão na operação de reservatórios. **Rbrh: revista brasileira de recursos hidricos**. Porto Alegre, RS. Vol. 13, n. 1 (jan./mar. 2008), p. 197-213, 2008.

BREIMAN, L. RandomForests. **MachineLearning**, v. 45, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Acesso em: 25 mar. 2023.

BUENO, J. C. M. *et al.* Fases do Enos e características da precipitação pluvial e temperatura na região de Guarapuava, Sul do Brasil. **Research, Society and Development**, v. 9, n. 5, p. e74953117-e74953117, 2020.

CAMPOS, J. N. B. VULNERABILIDADES HIDROLÓGICAS DO SEMI-ÁRIDO ÀS SECAS. **Planejamento e Políticas Públicas**, [S. l.], n. 16, 2022. Disponível em: [//www.ipea.gov.br/ppp/index.php/PPP/article/view/120](http://www.ipea.gov.br/ppp/index.php/PPP/article/view/120). Acesso em: 18 nov. 2023.

NOVAIS, R. P. de.; CARVALHO JÚNIOR, A. P. de.; OLIVEIRA, M. A. de. A Perenização de rios através da construção de açudes para o combate à seca no semiárido nordestino. **Geopauta**, [S. l.], v. 6, p. e9401, 2023. DOI: 10.22481/rg.v6.e2022.e9401.

Disponível em: <https://periodicos2.uesb.br/index.php/geo/article/view/9401>. Acesso em: 4 dez. 2023.

CARVALHO, T. M. N. **Water demand modelling using machine learning techniques**. Dissertação (Mestre) em Recursos Hídricos, Departamento de Engenharia Hidráulica e Ambiental, Universidade Federal do Ceará, Fortaleza, 2019.

CASAGRANDE, M. H. **Comparação de métodos de estimação para problemas com colinearidade e/ou alta dimensionalidade ($p > n$)**. 2016. Dissertação (Mestrado em Estatística) – Universidade Federal de São Carlos, São Carlos, 2016. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/7954>.

CASTRO, P. A. D. de. *et al.* Improving a Pittsburgh learnt fuzzy rule base using feature subset selection. In: **Fourth International Conference on Hybrid Intelligent Systems (HIS'04)**. IEEE, p. 180-185, 2004.

CERQUEIRA, E. O. de *et al.* Redes neurais e suas aplicações em calibração multivariada. **Química Nova**, v. 24, p. 864-873, 2001.

CHEIN, F. **Uma aplicação de redes neurais recorrentes do tipo lstm à previsão dos preços de curto prazo do mercado de energia elétrica brasileiro**. Brasília: Escola Nacional de Administração Pública (Enap), 2019. 76 p.

CHOLLET, F. *et al.* Keras: The Python Deep Learning API. 2015. URL: <https://keras.io> [accessed 2020-01-31], 2020.

CLARK, O. A. C.; CASTRO, A. A. A pesquisa. **Pesquisa Odontológica Brasileira**, v. 17, p. 67-69, 2003.

COPELAND, A. H. A “reasonable” social welfare function. **Seminar on Application of Mathematics to the Social Sciences**. University of Michigan. 1951.

COSTA, J. A.; SILVA, D. F. da. Distribuição espaço-temporal do Índice de anomalia de chuva para o Estado do Ceará (Distribution space-temporal of rain anomaly index for the Ceará State). **Revista brasileira de geografia física**, [S. l.], v. 10, n. 4, p. 1002-1013, 2017.

DASTILE, X.; CELIK, T.; POTSANE, M. Statistical and machine learning models in credit scoring: A systematic literature survey. **Applied Soft Computing**, v. 91, p. 106263, 2020.

DAWSON, C. W. *et al.* Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. **Hydrology and Earth System Sciences**, v. 6, n. 4, p. 619-626, 2002.

DAWSON, C. W.; WILBY, R. L. Hydrological modelling using artificial neural networks. **Progress in physical Geography**, v. 25, n. 1, p. 80-108, 2001.

DI, Yahui et al. Developing machine learning tools for long-lead heavy precipitation prediction with multi-sensor data. In: **2015 IEEE 12th International Conference on Networking, Sensing and Control**. IEEE, 2015. p. 63-68.

DIAS, T. L.; CATALDI, M.; FERREIRA, V. H. Application of neural networks techniques and atmospheric modeling to elaborate inflow forecasts in the Rio Grande Basin (MG), Brazil. **Engenharia Sanitaria e Ambiental**, v. 22, n. 1, p. 169-178, 2017.

EARTH SYSTEM RESEARCH LABORATORY. *Climate Indices: Monthly Atmospheric and Ocean Time-Series*. Broadway: ESRL, 2020.

EVANGELISTA, F. I. S. **Previsão climática sazonal para a região hidrográfica dos sertões de Crateús utilizando redes neurais artificiais**. 2019. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Civil) - Campus de Crateús, Universidade Federal do Ceará, Crateús, 2019.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. de. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011.

FERNANDES, W. S.; SILVEIRA, C. S. **Regularização de vazão a longo prazo para o Nordeste Brasileiro**. In: ADAPTA: Gestão adaptativa do risco climático de seca/ organizado por Francisco de Assis de Souza Filho *et al.* Fortaleza: Expressão Gráfica e Editora, 2019.

FERREIRA, A. G.; MELLO, N. G. da S. Principais sistemas atmosféricos atuantes sobre a Região Nordeste do Brasil e a influência dos Oceanos Pacífico e Atlântico no clima da região. **Revista Brasileira de Climatologia**, Curitiba, v. 1, n.1. p. 15-26, 2005.

FIORIN, V. *et al.* Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares. **Revista Brasileira de Ensino de Física**, v. 33, p. 01-20, 2011.

FURTADO, M. I. V. Redes neurais artificiais: uma abordagem para sala de aula. **Ponta Grossa, PR. Atena Editora**, p. 19, 2019.

GHARAT, S. **O quê, por que e qual?? Funções de Ativação**. 2019. Disponível em:<<https://medium.com/@snaily16/what-why-and-which-activation-functions-b2bf748c0441>>. Acesso em: 13 de novembro de 2019.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 1991.

GOMES, D. T. Redes neurais recorrentes para previsão de séries temporais de memórias curta e longa. **Master's thesis, Department of Statistics, Campinas State University, Campinas, Brazil**, p. 153, 2005.

GREFF, K. *et al.* LSTM: A search space odyssey. **IEEE transactions on neural networks and learning systems**, v. 28, n. 10, p. 2222-2232, 2016.

GUARNIERI, R. A. **Emprego de redes neurais artificiais e regressão linear múltipla no refinamento das previsões de radiação solar do modelo ETA**. Dissertação (Mestre) do Curso de Pós-Graduação em Meteorologia. INPE, São José dos Campos, 2006.

HAYKIN, S. **Neural networks and learning machines, 3/E**. Pearson Education India, 2009.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

HOFFMANN, R. **Análise de regressão: uma introdução à econometria**. 5^o. ed. Piracicaba: O autor, 2016. 393 p. DOI: 10.11606/9788592105709. Disponível em: https://www.esalq.usp.br/biblioteca/sites/default/files/Analise_Regress%C3%A3o.pdf. Acesso em: 5 dez. 2023.

HONORATO, A. G. S. M.; SILVA, G. B. L.; SANTOS, C. A. G. Previsão de Vazão Diário em uma Bacia Pequena Utilizando Modelos Híbridos de Transformadas Wavelet e Redes Neurais Artificiais. **XXII Simpósio Brasileiro de Recursos Hídricos**. Florianópolis-SC, 2017.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Brasileiro de 2022**. Rio de Janeiro: IBGE, 2023.

INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ - IPECE. **Ceará em mapas: Caracterização Territorial**. Fortaleza: IPECE 2021.

INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ – IPECE. **Perfil Municipal 2017 Crateús**. Fortaleza: IPECE, 2018.

JAMES, G. *et al.* **An introduction to statistical learning**. New York: springer, 2013.

KAPLAN, A.; CANE, M.; KUSHNIR, Y.; CLEMENT, A.; BLUMENTHAL, M.; RAJAGOPALAN, B. Analyses of global sea surface temperature 1856-1991. **Journal of Geophys Research**, v. 103, n. C9, p. 18,567-18,589, 1998.

KNIGHT, J. R.; FOLLAND, C. K.; SCAIFE, A. A. Climate impacts of the Atlantic multidecadal oscillation. **Geophysical Research Letters**, v. 33, n. 17, p. 2–5, 2006.

LEE, H. D. **Seleção de atributos importantes para a extração de conhecimento de bases de dados**. 2005. 154 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação - ICMC-USP, Recife, 2005.

LOPES, J.E.G.; BRAGA JR., B.P.F.; CONEJO, J.G.L. (1981). Simulação Hidrológica: Aplicações de um Modelo Simplificado. In: **Anais do III Simpósio Brasileiro de Recursos Hídricos**, v.2, 42-62, Fortaleza, 1981.

LORENZI, A. *et al.* Artificial neural networks application to predict bond steel-concrete in pull-out tests. **Revista IBRACON de Estruturas e Materiais**, v. 10, p. 1051-1074, 2017.

LUO, X. *et al.* A hybrid support vector regression framework for streamflow forecast. **Journal of Hydrology**, v. 568, p. 184-193, 2019.

MACEDO, J. A. *et al.* An intelligent decision support system to investment in the stock market. **IEEE Latin America Transactions**, v. 11, n. 2, p. 812-819, 2013.

MACHADO, F. W. **Modelagem Chuva-Vazão Mensal Utilizado Redes Neurais Artificiais**. Dissertação (Mestrado em Engenharia de Recursos Hídricos e Ambiental) - Universidade Federal do Paraná, Curitiba-PB, 2005.

MALFATTI, M. G. L.; CARDOSO, A. de O.; HAMBURGER, D. S. Modelo empírico linear para previsão de vazão de rios na usina hidrelétrica de Itaipu–Bacia do Rio Paraná. **Revista Brasileira de Meteorologia**, v. 33, p. 257-268, 2018.

MANTUA, N. J. *et al.* A Pacific interdecadal climate oscillation with impacts on salmon production. **Bulletin of the American Meteorological Society**, v. 78, n. 6, p. 1069-1080, 1997.

MARENGO, J A.; VALVERDE, M. C. Caracterização do clima no Século XX e Cenário de Mudanças de clima para o Brasil no Século XXI usando os modelos do IPCC-AR4. **Revista Multiciência**, Campinas – SP, v. 8, n. 1, p. 5-28, 2007.

MARINHO, F. P.; NETO, A. RR; ROCHA, P. AC. Previsao de Irradiância Solar de Curto Prazo Utilizando Modelo de Envelopes para os Preditores. In: **Congresso Brasileiro de Automática-CBA**. 2022.

MCKINNEY, Wes *et al.* pandas: a foundational Python library for data analysis and statistics. **Python for high performance and scientific computing**, v. 14, n. 9, p. 1-9, 2011.

MELO, B. V. de. **Automatização na decisão de negociação de ativos na bolsa de valores**. Orientador: Prof. Dr. Daniel Cruz Cavalieri. 2023. 78 p. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Controle e Automação) - Instituto Federal do Espírito Santo, Serra, 2023.

MENG, J. *et al.* Protein function prediction based on data fusion and functional interrelationship. **Mathematical biosciences**, v. 274, p. 25-32, 2016.

MITCHELL, T. M. **Machine Learning**. [S. l.]: McGraw-Hill Science/Engineering/Math, 1997. 432 p.

MONARD, M. C. *et al.* **Uma introdução ao aprendizado simbólico de máquina por exemplos**, notas Didáticas do ICMC-USP-São Paulo. 1997.

MONTEIRO, J. B.; ZANELLA, M. E.; PINHEIRO, D. R. de C. A contribuição da técnica dos quantis na identificação de extremos de chuva e de uma metodologia para detectar situações de desastre natural no semiárido cearense. **Revista Geografias**, [S. l.], v. 17, n. 2, p. 1–16, 2022. DOI: 10.35699/2237-549X.2021.36790. Disponível em: <https://periodicos.ufmg.br/index.php/geografias/article/view/36790>. Acesso em: 4 dez. 2023.

MORIASI, D. N. *et al.* Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. **American Society of Agricultural and Biological Engineers**, v. 50, n. 3, p. 885-900, 2007.

MOURA, A. D.; SHUKLA, J. On the dynamics of droughts in northeast Brazil: Observations, theory and numerical experiments with a general circulation model. **Journal of the atmospheric sciences**, v. 38, n. 12, p. 2653-2675, 1981.

NOBRE, P. **As origens das águas no Nordeste**: A questão da água no Nordeste. *In*: Centro de Gestão e Estudos Estratégicos, Agência Nacional de Águas. – Brasília, DF: CGEE, 2012.

NOBRE, P; SHUKLA, J. Variations of sea surface temperature, wind stress, and rainfall over the tropical Atlantic and South America. **Journal of climate**, v. 9, n. 10, p. 2464-2479, 1996.

OLAH, C. **Understanding lstm networks**. [S. l.], 2015. Disponível em: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Acesso em: 6 dez. 2023.

OLIVEIRA, B. A. S. *et al.* Avaliação de uma rede neural artificial como estimador temporal pluviométrico no sistema de abastecimento cantareira. **Revista de Informática Aplicada**, v. 14, n. 1, 2018.

OLIVEIRA, W.; CUNHA, L.; PAIVA, I. A influência da Oscilação do Atlântico Norte nos caudais dos rios Vouga e Mondego – relevância na manifestação de cheias e inundações. **Revista de Geografia e Ordenamento do Território**, n. 12, p. 255-279, 2017.

PAPACHARALAMPOUS, G.; TYRALIS, H.; KOUTSOYIANNIS, D. Predictability of monthly temperature and precipitation using automatic time series forecasting methods. **Acta Geophysica**, v. 66, p. 807-831, 2018.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, v. 12, p. 2825-2830, 2011.

RODRIGUES, B. D.; COUTINHO, M. D. L.; SAKAMOTO, M. S.; JACINTO, L. V. UMA ANÁLISE SOBRE AS CHUVAS NO CEARÁ BASEADA NOS EVENTOS DE EL NIÑO, LA NIÑA E NO DIPOLO DO SERVAIN DURANTE A ESTAÇÃO CHUVOSA. **Revista Brasileira de Climatologia**, [S. l.], v. 28, p. 507–519, 2021. Disponível em: <https://ojs.ufgd.edu.br/index.php/rbclima/article/view/14723>. Acesso em: 4 dez. 2023.

ROSA, C. S; ALEXANDRE, A. M. B. Análise de funções objetivo no contexto da determinação do regime hidrológico em bacia do semiárido nordestino. **XXIV SIMPÓSIO**

BRASILEIRO DE RECURSOS HÍDRICOS, Belo Horizonte - MG, 2021. Disponível em: <https://anais.abrhidro.org.br/job.php?Job=13073>. Acesso em: 4 dez. 2023.

ROSA, C. S.; ALEXANDRE, A. M. B.; CID, D. A. C. Operação de reservatórios do Sistema de Abastecimento dos Municípios da Região Hidrográfica dos Sertões de Crateús, Estado do Ceará. **XXIV SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS**, Belo Horizonte - MG, 2021. Disponível em: <http://www.repositorio.ufc.br/handle/riufc/63542>. Acesso em: 4 dez. 2023.

SANTOS, A. L. L. **Previsão de Vazão Afluente da UHE-Tucuruí por Redes Neurais Recorrentes LSTM**. Orientador: Prof. Dr. Raphael Barros Teixeira. 2022. 44 p. Trabalho de Conclusão de Curso (Bacharelado em engenharia elétrica) - Universidade Federal do Pará, [S. l.], 2022.

SANTOS, C. A. G.; SILVA, G. B. L. da. Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models. **Hydrological Sciences Journal**, v. 59, n. 2, p. 312-324, 2014.

SANTOS, G. **Uma aplicação de redes neurais recorrentes do tipo lstm à previsão dos preços de curto prazo do mercado de energia elétrica brasileiro**. Orientador: Prof. Dr. Afonso de Campos Pinto. 2019. 76 p. Tese (Mestrado em Economia.) - Escola de Economia de São Paulo da Fundação Getulio Vargas, São Paulo, 2019.

SANTOS, M. S.; LIMA, C. H. R.; MEDIERO, L. Modelo de cheias anuais baseado em mecanismos geradores em um contexto não-estacionário na bacia do rio paraguai. **I Encontro Nacional de Desastres**, Porto Alegre - RS, 2018

SECRETARIA DOS RECURSOS HÍDRICOS (CE). Parecer Técnico 05 de julho de 2023. **A BARRAGEM FRONTEIRAS E A RESILIÊNCIA HÍDRICA DA REGIÃO HIDROGRÁFICA DOS SERTÕES DE CRATEÚS**, Fortaleza - CE, 2023.

SECRETARIA DOS RECURSOS HÍDRICOS (Ceará). COMPANHIA DE GESTÃO DOS RECURSOS HÍDRICOS. **DIAGNÓSTICO DA REGIÃO HIDROGRÁFICA DOS SERTÕES DE CRATEÚS**. [S. l.: s. n.], 2021. Disponível em: https://portal.cogerh.com.br/wp-content/uploads/2021/07/DIAGN%C3%93STICO_RHSC_corrigeo_FINAL.pdf. Acesso em: 6 dez. 2023.

SHEYKHMUSA, M. *et al.* Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 13, p. 6308-6325, 2020.

SIAMI-NAMINI, S.; NAMIN, A. S. Forecasting economics and financial time series: ARIMA vs. LSTM. **arXiv preprint arXiv:1803.06386**, 2018.

SILVA, A. L. R. da. **Seleção de atributos para apoio ao diagnóstico do câncer de mama usando imagens termográficas, algoritmos genéticos e otimização por enxame de**

partículas. 2019. 62 p. Dissertação (Mestrado em Engenharia Biomédica) - Centro de Tecnologia e Geociências da Universidade Federal de Pernambuco, Recife, 2019.

SILVA, D. F. da. Influência Interdecadal (ODP e OMA) nas Cotas do Rio São Francisco. **Revista Brasileira de Geografia Física**, v. 6, n. 06, p. 1529-1538, 2013.

SILVA, D. F. da.; GALVÍNCIO, J. D. Estudo da Influência da Oscilação Decadal do Pacífico no Nordeste do Brasil (Study of the Influence of Pacific Decadal Oscillation in Northeast of Brazil). **Revista Brasileira de Geografia Física**, v. 4, n. 4, p. 665-676, 2011.

SILVA, E. L. da; MENEZES, Estera Muszkat. Metodologia da pesquisa e elaboração de dissertação. **UFSC, Florianópolis, 4a. edição**, v. 123, 2005.

SILVA, F. F.; SANTOS, F. de A. dos; SANTOS, J. M. dos. Índice de anomalia de chuva (IAC) aplicado ao estudo das precipitações no município de caridade, ceará, brasil. **Revista Brasileira de Climatologia**, [S. l.], v. 27, p. 426-442, 2021.

SILVA, G. de A.; FARIAS, C. W. L. de A. Análise espaço-temporal da vegetação no semiárido do Nordeste brasileiro utilizando parâmetros biofísicos. **Revista Semiárido De Visu**, [S. l.], v. 9, n. 3, p. 195-203, 2021.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes Neurais Artificiais Para Engenharia E Ciências Aplicadas: Curso Prático**. 1ª Edição. Editora Artliber, 2010.

SIQUEIRA A. H. B. **Variabilidade do clima da América do Sul e sua relação com os índices oceânicos e atmosféricos**. Dissertação (Mestre) em meteorologia, Universidade Federal de Alagoas, Instituto de Ciências Atmosféricas, Maceió, 2012.

SOUSA, W. A. **Disponibilidade hídrica da região hidrográfica dos sertões de Crateús: a regularização de vazões por reservatórios estratégicos**. 2022. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Civil) - Campus de Crateús, Universidade Federal do Ceará, Crateús, 2022. Disponível em: <http://www.repositorio.ufc.br/handle/riufc/65666>. Acesso em: 06/11/2023.

SOUZA FILHO, F. A.; LALL, U. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. **Water Resources Research**, v. 39, n. 11, 2003.

SOUZA NETO, P. F. de.; SILVA, D. F. da.; ALMEIDA, H. R. R. de C. Análise da variabilidade climática dos oceanos Atlântico e Pacífico. **Revista Brasileira de Geografia Física**, [S. l.], v. 14, n. 04, p. 1861-1879, 2021.

SRINIVASAN, D.; LIEW, A. C.; CHANG, C. S. A neural network short-term load forecaster. **Electric Power Systems Research**, v. 28, n. 3, p. 227-234, 1994.

SRIVASTAVA, N. *et al.* Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, v. 15, n. 1, p. 1929-1958, 2014.

STOKELJ, T.; PARAVAN, D.; GOLOB, R. Enhanced artificial neural network inflow forecasting algorithm for run-of-river hydropower plants. **Journal of Water Resources Planning and Management**, v. 128, n. 6, p. 415-423, 2002.

TANG, Z.; FISHWICK, P. A. Feedforward neural nets as models for time series forecasting. **ORSA journal on computing**, v. 5, n. 4, p. 374-385, 1993.

TRIGO, R.; OSBORN, T.; CORTE-REAL, J. Influência da Oscilação do Atlântico Norte no clima do continente europeu e no caudal dos rios ibéricos atlânticos. **Finisterra**, [S. l.], v. 37, n. 73, 2002. DOI: 10.18055/Finis1606. Disponível em: <https://revistas.rcaap.pt/finisterra/article/view/1606>. Acesso em: 5 dez. 2023.

UVO, C.; BERNDTSSON, R.; Regionalization and spatial properties of Ceará State rainfall in northeast Brazil. **Journal of Geophysical Research: Atmospheres**, v. 101, n. D2, p. 4221-4233, 1996.

VASCO, L. P. **Um estudo de redes neurais recorrentes no contexto de previsões no mercado financeiro**. 2020. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) – Universidade Federal de São Carlos, São Carlos, 2020.

ALEXANDRE, A. M. B.; BATISTA, T. L.; SOUSA, W. A. Influência das normais climatológicas de evaporação na regularização de vazões em reservatórios superficiais. **XVI SRHNe - Simpósio de Recursos Hídricos do Nordeste e 15º SILUSBA - Simpósio de Hidráulica e Recursos Hídricos dos Países de Língua Portuguesa: Mudanças Climáticas e Águas Continentais e costeiras**, Caruaru - PE, 2022.

XU, Y. *et al.* Multi-instance multi-label distance metric learning for genome-wide protein function prediction. **Computational biology and chemistry**, v. 63, p. 30-40, 2016.

ZANELLA, L. C. H. **Metodologia de Pesquisa**. 2º. ed. rev. e atual. Florianópolis - SC: Departamento de Ciências da Administração/ UFSC, 2011. 136 p. ISBN 978-85-7988-111-3.

ZAREMBA, W.; SUTSKEVER, I.; VINYALS, O. Recurrent neural network regularization. **arXiv preprint arXiv:1409.2329**, 2014.

ZHENG, H.; YUAN, J.; CHEN, Long. Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. **Energies**, v. 10, n. 8, p. 1168, 2017.