UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS DE ANDRADE AMARAL

APLICANDO REDE NEURAL PARA REALIZAR PREDIÇÕES DE NÍVEIS DO RIO TAPAJÓS NA REGIÃO OESTE DO PARÁ

Santarém – PA
2024

LUCAS DE ANDRADE AMARAL

APLICANDO REDE NEURAL PARA REALIZAR PREDIÇÕES DE NÍVEIS DO RIO TAPAJÓS NA REGIÃO OESTE DO PARÁ

Trabalho de conclusão de curso - TCC apresentado ao programa de computação para obtenção do grau de Bacharelado em Ciência da Computação do Instituto de Engenharia e Geociências - IEG na Universidade Federal do Oeste do Pará.

Orientador(a): Dr. Marcelino Silva da Silva.

Santarém – PA
2024

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/UFOPA

A485a Amaral, Lucas de Andrade

Aplicando rede neural para realizar predições de níveis do Rio Tapajós na região Oeste do Pará. / Lucas de Andrade Amaral. - Santarém, 2024.

72 p. : il.

Inclui bibliografias.

Orientador: Marcelino Silva da Silva.

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências, Bacharelado em Ciência da Computação.

1. Rede Neural. 2. LSTM. 3. Rio Tapajós. 4. Predição. 5. Cotas. I. Silva, Marcelino Silva da, *orient*. II. Título.

CDD: 23 ed. 006.32

Bibliotecária - Documentalista: Cátia Alvarez – CRB/2 843

LUCAS DE ANDRADE AMARAL

APLICANDO REDE NEURAL PARA REALIZAR PREDIÇÕES DE NÍVEIS DO RIO TAPAJÓS NA REGIÃO OESTE DO PARÁ

Trabalho de conclusão de curso - TCC apresentado ao programa de computação para obtenção do grau de Bacharelado em Ciência da Computação do Instituto de Engenharia e Geociências - IEG na Universidade Federal do Oeste do Pará.

Conceito: 8,0
Data de Aprovação: 28/10/2024

Documento assinado digitalmente
gov.br
MARCELINO SILVA DA SILVA
Data: 13/11/2024 15:33:50-0300
Verifique em https://validar.iti.gov.br

_____
Dr. Marcelino Silva da Silva - Orientador
Orientador - Universidade Federal do Oeste do Pará (UFOPA)

Documento assinado digitalmente
gov.br
FABIO MANOEL FRANCA LOBATO
Data: 13/11/2024 11:24:13-0300
Verifique em https://validar.iti.gov.br

_____
Dr. Fábio Manoel França Lobato
Universidade Federal do Oeste do Pará (UFOPA)

Documento assinado digitalmente
gov.br
LIVIANE PONTE REGO
Data: 13/11/2024 15:22:30-0300
Verifique em https://validar.iti.gov.br

_____
Drª. Liviane Ponte Rêgo
Universidade Federal do Oeste do Pará (UFOPA)

À minha mãe, Sueli, e à minha avó, Maria José, inspirações de força e resiliência; e aos meus tios, Wander (in memoriam) e Ádria, que foram referência e motivação para os meus estudos.

AGRADECIMENTOS

Inicialmente devo agradecer a minha família, que tem sido a base de tudo na minha vida, começando pela minha mãe Sueli, minha avó Maria José e a minha irmãzinha Maria Rebeca, e um agradecimento especial aos meus tios, Wander e Ádria, que me acolheram nesta cidade para que eu conseguisse realizar o sonho de uma graduação, sem o incentivo deles, eu poderia ter tomado um outro caminho, então vocês mudaram o rumo da minha vida e por consequência o meu futuro.

Agradeço ao meu orientador, Marcelino, pela confiança depositada em mim na escolha do tema desta pesquisa e pelo valioso direcionamento oferecido ao longo de todo o processo, até a finalização deste trabalho.

A Universidade Federal do Oeste do Pará pela oportunidade da graduação de forma gratuita, pela educação de qualidade e por todas as experiências enriquecedoras de projetos de pesquisa que me propuseram alcançar ao meu estado profissional atual.

RESUMO

O objetivo deste trabalho foi utilizar uma rede neural LSTM para realizar predições dos níveis de cota do rio Tapajós, baseando-se em dados do acervo hidroclimatológico da Agência Nacional de Águas (ANA). Foram utilizados dados dos anos de 1999 a 2023 de três estações de hidro-telemetria estrategicamente posicionadas ao longo do rio, essas estações fornecem informações hidrológicas. A pesquisa adotou a metodologia KDD (Knowledge Discovery in Databases) para auxiliar no processamento e na organização dos dados, visando aprimorar os resultados e as interpretações das análises. O modelo LSTM foi construído e treinado para gerar predições mensais dos níveis de cota. A normalização Min-Max foi realizada para padronizar os dados de entrada facilitando o modelo a convergir o processo de validação com as predições. O modelo foi configurado para utilizar uma janela de observação de 30 dias. As métricas MAE, MSE e RMSE foram empregadas para avaliar o desempenho do modelo, como resultado, o modelo LSTM apresentou um desempenho satisfatório na predição dos níveis de cota do rio Tapajós. As métricas de avaliação utilizadas mostraram valores consistentes, considerando o melhor resultado dos 3 conjuntos de dados, o erro médio absoluto (MAE) foi de 0.035, o erro quadrático médio (MSE) atingiu 0.002, e a raiz do erro quadrático médio (RMSE) foi de 0.046 em uma sequência temporal contendo 1333 amostras, indicando boa precisão nas previsões. O LSTM foi capaz de capturar as tendências e variações sazonais presentes nos dados históricos, com base nisso foi realizada a técnica de predição auto regressiva para os 5 primeiros meses de 2024 para gerar dados de predições e compará-los com os dados reais disponíveis desse período.

Palavras-chave: Rede Neural. LSTM. Rio Tapajós. Predição. Cotas.

ABSTRACT

The objective of this study was to use an LSTM neural network to predict the water level of the Tapajós River, based on hydroclimatological data from the National Water Agency (ANA). Data from the years 1999 to 2023 were used from three hydro-telemetry stations strategically positioned along the river, providing hydrological information. The research adopted the Knowledge Discovery in Databases (KDD) methodology to assist in data processing and organization, aiming to improve the results and interpretations of the analyses. The LSTM model was built and trained to generate monthly predictions of water levels. Min-Max normalization was applied to standardize the input data, facilitating model convergence during validation and prediction processes. The model was configured to use a 30-day observation window. The MAE, MSE, and RMSE metrics were used to evaluate the model's performance; as a result, the LSTM model showed satisfactory performance in predicting the Tapajós River's water levels. The evaluation metrics indicated consistent values; considering the best results from the three datasets, the mean absolute error (MAE) was 0.035, the mean squared error (MSE) reached 0.002, and the root mean squared error (RMSE) was 0.046, in a time series containing 1,333 samples, indicating good predictive accuracy. The LSTM was able to capture seasonal trends and variations present in historical data. Based on this, an autoregressive prediction technique was used for the first five months of 2024 to generate forecast data and compare them with actual data available for this period.

Keywords: Neural Network. LSTM. Tapajós River. Prediction. Quotas.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ANA Agência Nacional de Águas e Saneamento Básico
ANN Artificial Neural Network
CNN Convolutional Neural Network
CSV Comma-separated Values
DL Deep Learning
GAN Generative Adversarial networks
IA Inteligência Artificial
KDD Knowledge Discovery in Databases
LSTM Long Short-Term Memory
MAE Mean Absolute Error
ML Machine Learning
MSE Mean Squared Error
NN Neural Network
RNN Recurrent Neural Network
RHN Rede Hidrometeorológica Nacional
RMSE Root Mean Square Error
SNIRH Sistema Nacional de Informações sobre Recursos Hídricos

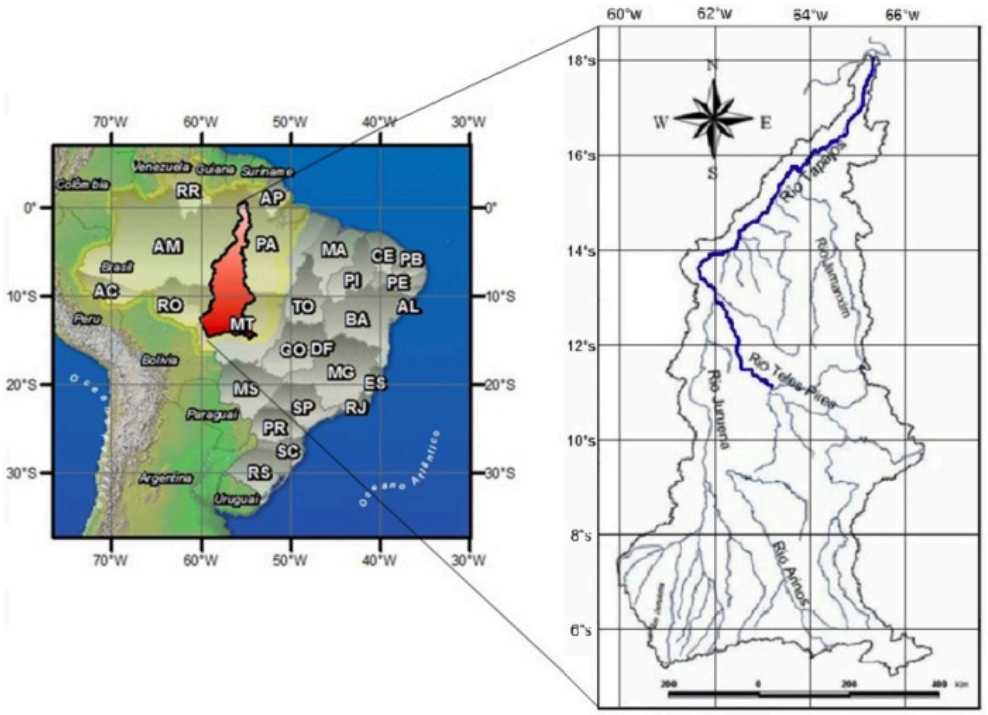
SUMÁRIO

1. INTRODUÇÃO

A região Oeste do Pará é uma mesorregião composta por 29 municípios, um dos seus principais rios é o Tapajós, rio pertencente a bacia hidrográfica do Tapajós (Figura 1), que se encontra à margem direita do rio Amazonas. A bacia abrange três estados Brasileiros, sendo: Pará 38%, Amazonas 3% e Mato Grosso 59%, tornando-a uma bacia federal, com extensão de aproximadamente 764.183 km².

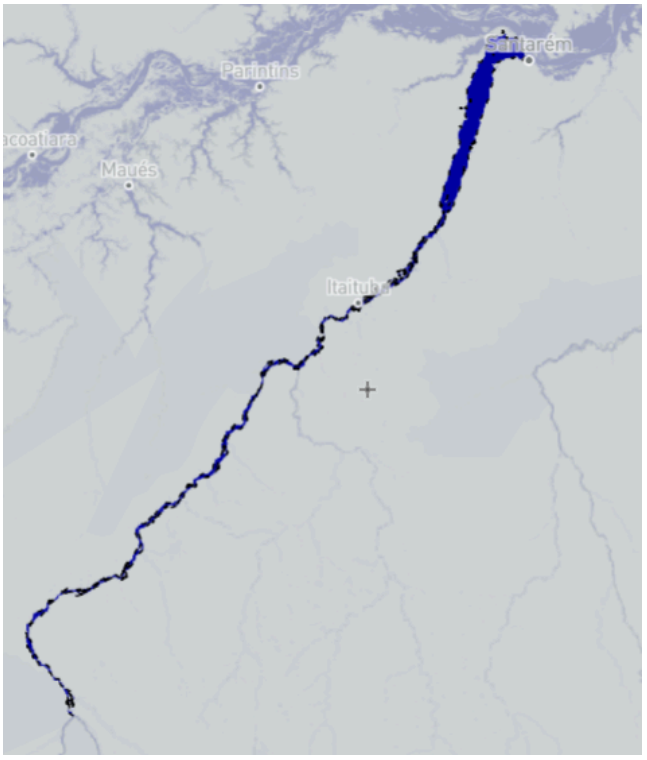
O rio Tapajós (Figura 2) possui uma extensão de 1.784 km e tem como características as águas claras. Seus afluentes são o rio São Manuel (também chamado de Teles Pires), e o rio Juruena, ambos localizados no estado do Mato Grosso, e sua foz se encontra com o rio Amazonas no município de Santarém, no Pará.

Figura 1 - Bacia Hidrogŕafica do Tapajós



Fonte:
https://www.researchgate.net/figure/Figura-1-Localizacao-da-bacia-do-rio-Tapajos-A-bacia-do-rio-Tapajos-esta-situada-nos_fig1_278609243

Figura 2 - Rio Tapajós



Fonte:
https://plataforma.agua.mapbiomas.org/water/-4.910625/-55.918598/5.8/brazil/naturalWaterMass/16731/naturalWaterMass/surface/2022/2022

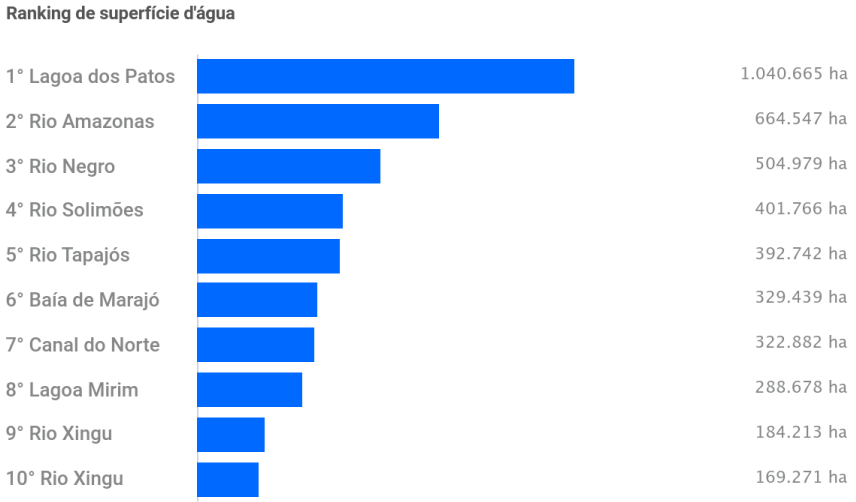
A precipitação média anual é de aproximadamente 2300 mm, existindo uma pronunciada estação seca de três a quatro meses. A estação chuvosa no alto Tapajós se inicia no final de setembro, enquanto que no baixo curso se inicia no final de dezembro ou janeiro. O pico das inundações nos cursos médio e alto do Tapajós normalmente ocorre em março. Próximo à boca do rio Tapajós, o nível do rio atinge seu nível mais alto normalmente em maio ou junho. Esta diferença nos períodos de alagamento de distintos trechos do rio acontece porque o nível das águas do Tapajós é controlado pelo rio Amazonas próximo à foz.

A flutuação anual do nível do Tapajós e dos rios Juruena e Teles Pires, dois de seus principais tributários, são em média de 4 a 5 metros. As diferenças entre os níveis máximo e mínimo do rio estão em torno de 8 a 9 metros.

A bacia hidrográfica do tapajós possui um grande potencial energético (hidroeletricidade), abastecimento de água para o consumo humano e dessedentação de animais. Sendo um exemplo da função social das águas como um bem de consumo final ou intermediário na quase totalidade das atividades humanas (SANTOS et al., 2010). Destacando-se a sua importância como a 5ª bacia

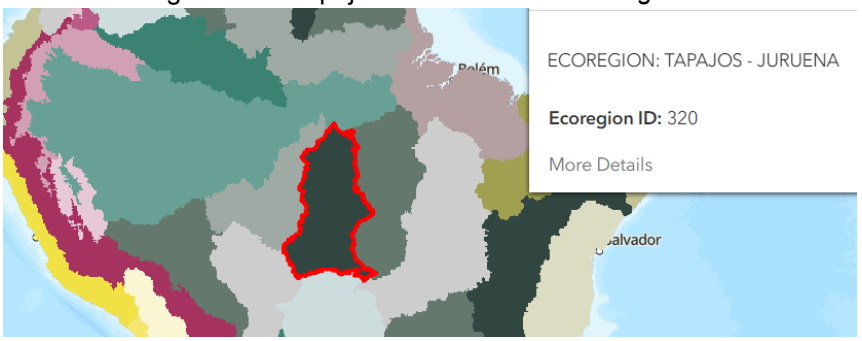
hidrográfica com maior superfície de água no Brasil (Figura 3 e 4), bacia essa que contribui para a manutenção dos ecossistemas amazônicos.

Figura 3 - Ranking de bacias hidrográficas



Fonte: https://plataforma.agua.mapbiomas.org/water/-4.910625/-55.918598/5.8/brazil/naturalWaterMass/16731/naturalWaterMass/surface/2022/2022

Figura 4 - Bacia Hidrografica do Tapajós no *Freshwater Ecoregions of the World* (FEOW)



Fonte: https://www.feow.org/ecoregions/details/320

De acordo com a ANA (2014), o conceito de cota refere-se à altura do nível da água em um ponto específico em relação a um ponto de referência, geralmente o nível do mar. As cotas são amplamente utilizadas no monitoramento de corpos d'água, como rios e lagos, sendo fundamentais para a gestão de recursos hídricos e a predição de eventos como enchentes.

A inteligência artificial tem se tornado uma área de grande relevância nos últimos anos e ela trouxe uma significativa contribuição ao permitir o desenvolvimento de algoritmos eficientes que auxiliam na tomada de decisões,

especialmente na mitigação de problemas complexos, podemos dizer que os desafios ambientais se encaixam nesse escopo. Os algoritmos de IA podem possibilitar novas formas eficientes de monitorar recursos naturais e prever eventos climáticos, fornecendo informações valiosas para a gestão sustentável do planeta.

Modelos avançados de aprendizado de máquina, como as redes neurais artificiais e redes neurais recorrentes são capazes de identificar padrões em grandes volumes de dados, aprimorando a precisão das análises em séries temporais, através da aplicação de metodologias eficazes para o tratamento e transformação de dados, é possível alcançar alta assertividade nas previsões relacionadas a variáveis ambientais. Esses recursos se tornam essenciais para análises detalhadas do comportamento de rios e outros sistemas naturais, independentemente das variáveis monitoradas.

1.1. Objetivos

1.1.1. Objetivo geral

Utilizar rede neural LSTM para fazer predições nos níveis de cota do rio Tapajós.

1.1.2. Objetivos específicos

- Desenvolver e treinar uma rede neural LSTM capaz de prever com precisão os níveis mensais de cota do rio Tapajós.
- Aplicar mineração de dados aos dados hidroclimatológicos disponíveis publicamente, gerando novos conhecimentos e compreensão sobre as dinâmicas hidrológicas do rio.
- Realizar uma revisão sistemática de publicações que utilizam técnicas de aprendizado de máquina no monitoramento de ambientes fluviais, identificando as abordagens mais eficazes.

1.2. Justificativa

Nos últimos anos, observa-se um aumento na frequência de eventos climáticos extremos, como enchentes e secas prolongadas, fenômenos que têm sido amplamente associados às mudanças climáticas globais (SILVA; PEREIRA, 2021).

O monitoramento contínuo dos níveis dos rios é essencial para a prevenção de desastres naturais e para a gestão eficiente dos recursos hídricos, fornecendo dados importantes para a tomada de decisões em situações de risco, como enchentes ou secas (AGÊNCIA NACIONAL DE ÁGUAS – ANA, 2014).

Por isso, a predição dos níveis de cota de rios é importante para a gestão de recursos hídricos, principalmente em regiões como a bacia do rio Tapajós, que desempenha um papel vital para o ecossistema, a economia local e a vida das comunidades ribeirinhas. A região Oeste do Pará, onde o rio Tapajós está localizado, depende diretamente da estabilidade dos níveis fluviais para atividades como agricultura, pesca, transporte e abastecimento de água.

Tradicionalmente, modelos hidrológicos clássicos têm sido amplamente utilizados para prever os níveis dos rios, mas esses métodos, em geral, porém se analisarmos as potencialidades de algoritmos de aprendizado de máquina podemos obter resultados mais precisos, dada a maneira de como esses algoritmos possuem capacidade de aprender padrões em dados. Nesse contexto, o uso de redes neurais LSTM apresenta uma abordagem promissora, pois esses modelos têm a capacidade de lidar com grandes volumes de dados e capturar padrões complexos em séries temporais, o que pode resultar em predições mais precisas e confiáveis.

Além disso, o uso de dados públicos, como os fornecidos pela Agência Nacional de Águas (ANA), e o processo de mineração de dados (KDD) permitem gerar novos conhecimentos a partir de informações já disponíveis, potencializando o impacto de estudos como este, que podem auxiliar na melhoria da gestão hídrica regional e nacional.

A escolha de uma rede neural LSTM justifica-se pela sua eficácia comprovada na modelagem de séries temporais, sobretudo em cenários de variabilidade hidrológica, o que faz dela uma ferramenta adequada para o monitoramento e a predição dos níveis.

1.3. Organização do trabalho

Este trabalho está dividido em 7 seções.

1. Introdução: Contextualiza o objeto de estudo, o problema a ser tratado e as contribuições dos algoritmos de redes neurais para predição em séries temporais. Também inclui o objetivo geral e os específicos desta pesquisa.

2. Revisão sistemática: Apresenta uma revisão prévia de trabalhos relacionados ao uso de algoritmos de machine learning aplicados ao monitoramento ambiental, com foco específico em séries temporais e predição de variáveis hidrológicas.

3. Fundamentação teórica: Uma descrição sobre os processos hidrológicos, Inteligência artificial e seus subcampos de pesquisa, o modelo de rede neural escolhido, além da descrição da metodologia KDD.

4. Metodologia: Contém todas as etapas para o desenvolvimento deste trabalho, desde a aquisição dos dados até a construção e treinamento do modelo LSTM para predições.

5. Resultados: Análises das métricas de avaliação, desempenho do modelo e gráficos dos resultados do treinamento.

6. Contribuições e proposta de continuação da pesquisa: Apresenta contribuições ao demonstrar a aplicação de redes neurais recorrentes no monitoramento hidrológico do rio Tapajós, com foco na predição de níveis de cota a partir de dados hidroclimatológicos. Além disso, propõe uma continuidade na pesquisa com a modelagem de um conjunto mais robusto de modelos de redes neurais, integrando variáveis adicionais, como precipitação, cota e vazão.

7. Referências: Contém todo o material bibliográfico além dos repositórios de código e as plataformas públicas de consulta para os dados.

2. REVISÃO SISTEMÁTICA

Para esta pesquisa, foi realizada uma revisão sistemática da literatura com o objetivo de identificar estudos que utilizam algoritmos de machine learning (ML) e deep learning (DL) no monitoramento ambiental de rios e ambientes fluviais. O processo de levantamento de trabalhos seguiu uma metodologia estruturada, que incluiu a formulação de perguntas de pesquisa, a definição de critérios de inclusão e exclusão, e a construção de uma search string apropriada para identificar os estudos relevantes em diferentes bases de dados científicas.

2.1. Perguntas de Pesquisa (Research Questions)

2.1.1. Pergunta Principal:

- Quais trabalhos utilizam algoritmos de Machine Learning e Deep Learning para estudos ambientais fluviais?

2.1.2. Perguntas Adicionais:

1. Quais são os principais algoritmos de ML e DL utilizados em estudos de ambientes fluviais?
2. Quais aspectos das dinâmicas fluviais são mais abordados por esses estudos?
3. Qual é a eficácia dos modelos de ML e DL na análise, previsão ou predição de mudanças nos ambientes fluviais?
4. Quais são os principais desafios e limitações enfrentados pelos estudos que utilizam ML e DL em análises fluviais?

A revisão foi conduzida seguindo protocolos padrões para garantir a qualidade e relevância dos estudos incluídos, desde a formulação de questões específicas até a seleção cuidadosa de artigos que atendessem aos critérios estabelecidos. Isso permitiu uma análise abrangente das abordagens e tecnologias utilizadas no campo da ciência ambiental fluvial, focando no uso de técnicas avançadas de aprendizado de máquina e aprendizado profundo.

2.1.3. Resultados

Após a definição das diretrizes, foi realizada uma análise quantitativa dos trabalhos catalogados nas três principais bases de dados (tabela 1), seguindo os protocolos estabelecidos anteriormente. O objetivo dessa análise foi obter uma visão abrangente e clara sobre os modelos mais utilizados nos estudos analisados, permitindo identificar tendências e padrões no uso de algoritmos de Machine Learning e Deep Learning aplicados ao monitoramento de ambientes fluviais (Figuras 5 a 7)

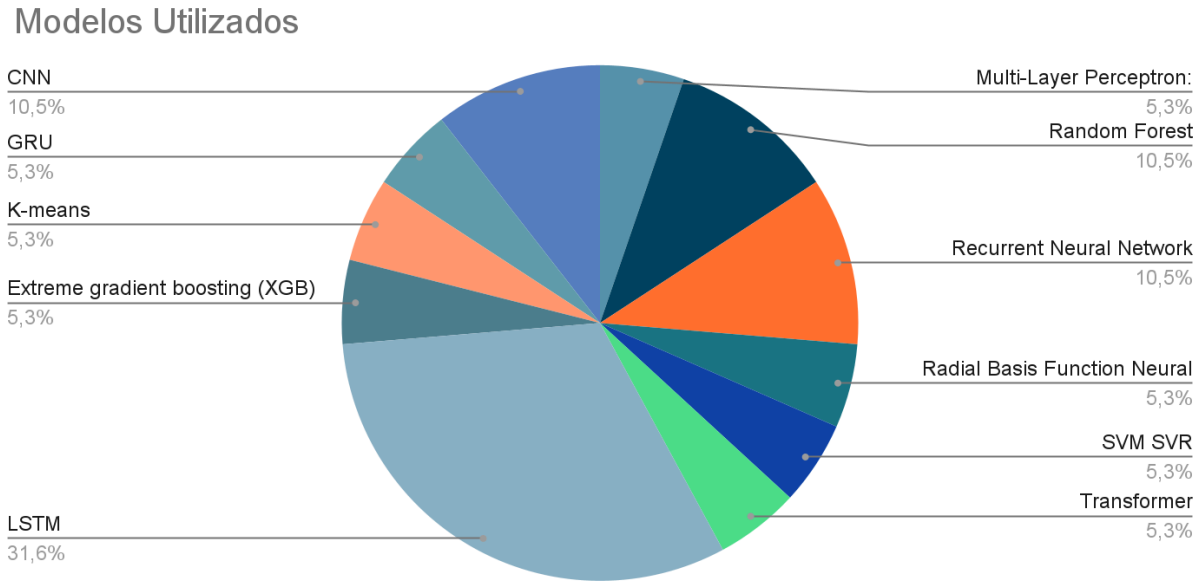
Tabela 1 - Trabalhos coletados da revisão sistemática

Base de dados	Resultados	Acesso público	Artigos coletados	Artigos selecionados	Nº total dos selecionados
IEEE Xplore	286	47	22	9	
Scielo	17	17	17	7	
ArXiv	23	23	16	8	
					24

Fonte: Autor

Base de dados IEEE Xplore

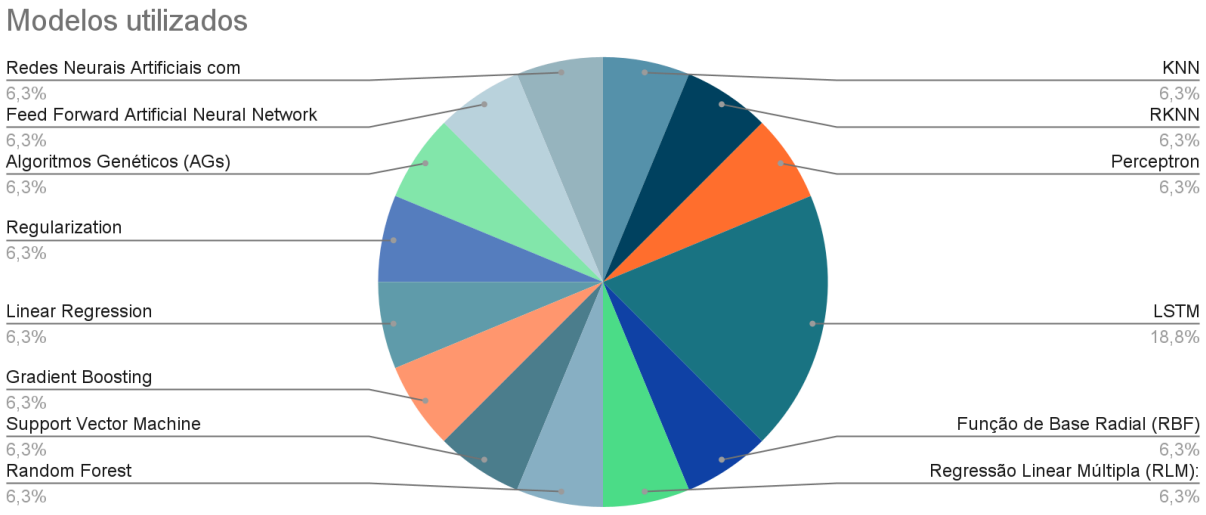
Figura 5 - Resultados IEEE Xplore



Fonte: Autor

Base de dados Scielo

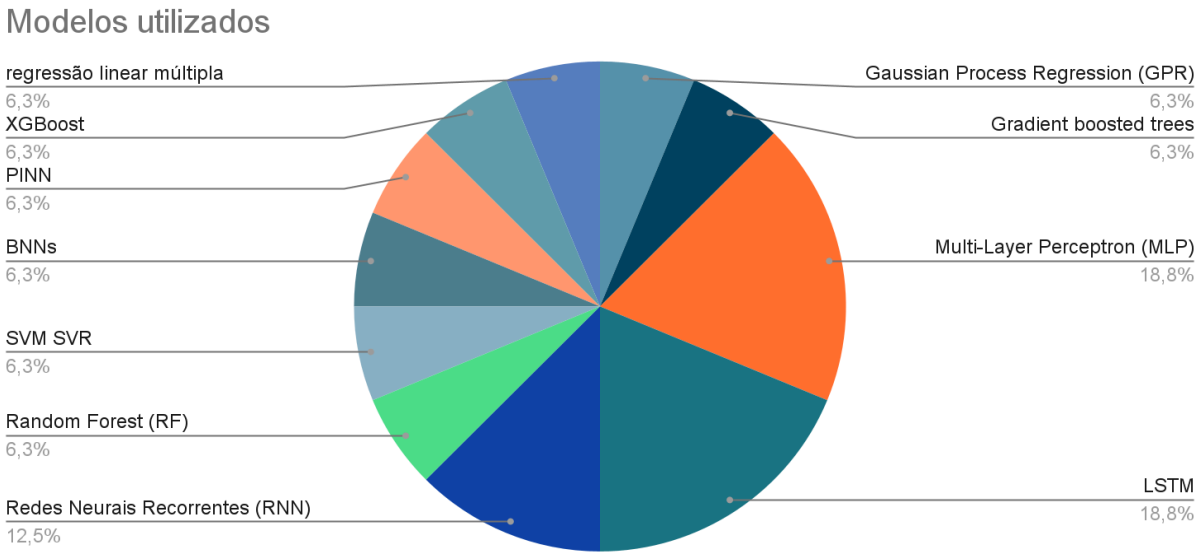
Figura 6 - Resultados Scielo



Fonte: Autor

Base de dados ArXiv

Figura 7 - Resultados ArXiv



Fonte: Autor

Foi observado que grande parte dos trabalhos de monitoramento utilizam redes neurais recorrentes, como LSTM, e redes feedforward, como multilayer perceptron, focam em análises de séries temporais e dados tabulados para monitoramento ambiental.

Esses estudos frequentemente exploram padrões e tendências em dados históricos, utilizando essas arquiteturas para prever comportamentos futuros em contextos de mudanças ambientais, variáveis hidrológicas e meteorológicas. Além disso, os resultados indicam que essas redes neurais são escolhidas devido à sua capacidade de lidar com dependências temporais e não linearidades presentes nos dados ambientais, permitindo uma modelagem mais precisa e robusta das condições fluviais ao longo do tempo.

A prevalência de dados tabulados nesses estudos reflete a natureza estruturada dos dados de monitoramento ambiental, facilitando a aplicação de técnicas de aprendizado de máquina para análise preditiva e tomada de decisão em cenários complexos de gestão e conservação ambiental.

Essa pesquisa foi essencial para a escolha do modelo de rede neural a ser utilizada na pesquisa do trabalho de predição de níveis de cota, os detalhamentos do modelo serão descritos nos capítulos subsequentes.

3. FUNDAMENTAÇÃO TEÓRICA

3.1. Hidrologia e as dinâmicas dos rios

A hidrologia é a ciência que estuda a ocorrência, distribuição e movimento da água na superfície terrestre, incluindo o ciclo hidrológico, os corpos d'água e a interação desses elementos com o meio ambiente. Nos rios, o comportamento dinâmico da água é influenciado por diversos fatores, como a precipitação, evapotranspiração, infiltração e escoamento superficial, que determinam o regime fluvial ao longo do tempo (TUCCI, 2009).

O ciclo hidrológico é o processo contínuo de movimentação da água entre a atmosfera, a superfície terrestre e os corpos d'água. Nos rios, esse ciclo é fundamental para entender como as variações de precipitação influenciam o comportamento dos níveis de água, conhecidos como cotas. A água que chega aos rios por meio das chuvas, derretimento de neve e outros processos naturais altera significativamente o volume e a vazão do curso d'água, levando à necessidade de monitoramento constante (LINSLEY; KOHLER; PAULHUS, 1982).

A cota de um rio refere-se à altura da superfície da água em um ponto específico do curso d'água em relação a um nível de referência, como o nível médio do mar ou um ponto de controle estabelecido por uma estação de monitoramento. Este conceito é essencial para a hidrologia, pois permite o acompanhamento das flutuações de nível do rio e a predição de eventos críticos, como enchentes e secas (AGÊNCIA NACIONAL DE ÁGUAS – ANA, 2014).

Cota normal: O nível do rio em condições de equilíbrio, ou seja, sem grandes variações devido a chuvas intensas ou estiagens severas.

Cota de alerta: Um nível que sinaliza uma aproximação de uma situação de risco, como o aumento repentino da vazão devido a chuvas fortes ou prolongadas.

Cota de inundação: Quando o nível do rio ultrapassa o limite crítico, atingindo áreas ribeirinhas e provocando inundações. Este dado é fundamental para alertar as autoridades e implementar medidas de prevenção (ANA, 2014).

A dinâmica dos rios está diretamente ligada à sua bacia hidrográfica, que é a área de captação natural da água da chuva, drenando-a para o curso principal. As bacias hidrográficas variam em tamanho e complexidade, influenciando diretamente o comportamento hidrológico dos rios. Fatores como a topografia, o tipo de solo e a vegetação da bacia afetam o escoamento superficial e o tempo de resposta das

chuvas no nível dos rios (TUCCI, 2009). Além disso, as ações humanas, como a construção de barragens, desmatamento e urbanização, alteram drasticamente a dinâmica natural dos rios, intensificando eventos como enchentes e erosões.

O monitoramento contínuo dos níveis dos rios e sua dinâmica é essencial para a gestão dos recursos hídricos, especialmente em regiões suscetíveis a eventos climáticos extremos, como a bacia do rio Tapajós. Este tipo de monitoramento possibilita a antecipação de situações de crise, como secas severas ou inundações, permitindo a implementação de medidas de mitigação e adaptação (FERNANDES; GUERRA, 2016). Além disso, os dados hidrológicos são essenciais para a gestão de infraestruturas, como barragens e reservatórios, que dependem do fluxo adequado de água para seu funcionamento seguro.

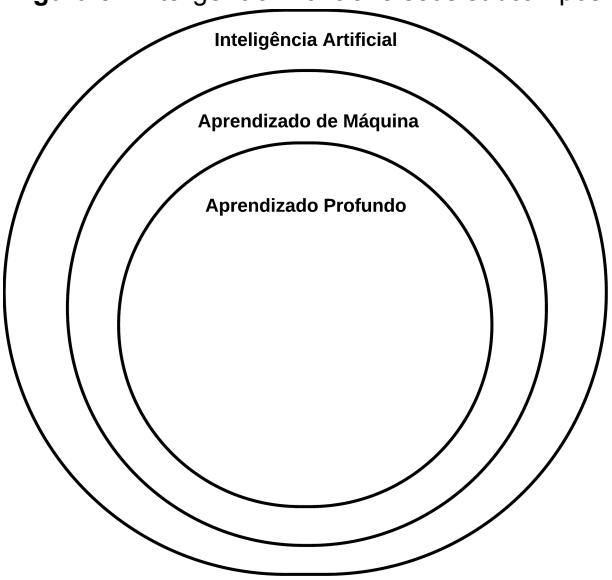
3.2. Aprendizado de máquina, aprendizado profundo e as redes neurais artificiais

O aprendizado de máquina (ML) é um grande subcampo pertencente a inteligência artificial (AI) (Figura 8), que se concentra em criar sistemas capazes de aprender e melhorar automaticamente a partir de experiências, sem serem explicitamente programados para isso.

Os algoritmos de aprendizado de máquina (ML) utilizam dados para identificar padrões e realizar tarefas como classificação, regressão, clustering e tomada de decisões. O aprendizado pode ser supervisionado, não supervisionado ou por reforço, dependendo da natureza dos dados e do problema a ser resolvido.

Dentro do aprendizado de máquina (ML), temos o Aprendizado Profundo (DL) um subcampo que abrange o uso de redes neurais artificiais com múltiplas camadas para modelar e aprender representações de dados em níveis progressivamente mais abstratos. Essa abordagem tem sido fundamental para avanços significativos em áreas como reconhecimento de imagem, processamento de linguagem natural, reconhecimento de voz, e muitos outros.

Figura 8 - Inteligência Artificial e seus subcampos



Fonte: Autor

As redes neurais artificiais (ANN) são modelos computacionais inspirados no funcionamento das redes neurais biológicas do cérebro humano, projetados para reconhecer padrões a partir de dados. Elas são compostas por diversas camadas que permitem que o modelo capture características complexas dos dados, trabalhando em tarefas de grande escala e complexidade.

Dentro do campo das redes neurais artificiais, podemos encontrar diversas arquiteturas e tipos (Figura 9). Dentre algumas destacadas, temos as redes transformer, redes neurais convolucionais, redes neurais generativas e as redes neurais recorrentes, cada uma com aplicações distintas e vantajosas para tipos específicos de problemas.

Redes Transformer: Essenciais em tarefas de processamento de linguagem natural, como tradução automática, geração de texto e análise de sentimentos. Elas capturam dependências entre palavras em uma frase, mesmo quando distantes, através do mecanismo de atenção, que permite que o modelo foque em diferentes partes do contexto para gerar respostas mais precisas e coerentes (Vaswani et al., 2017).

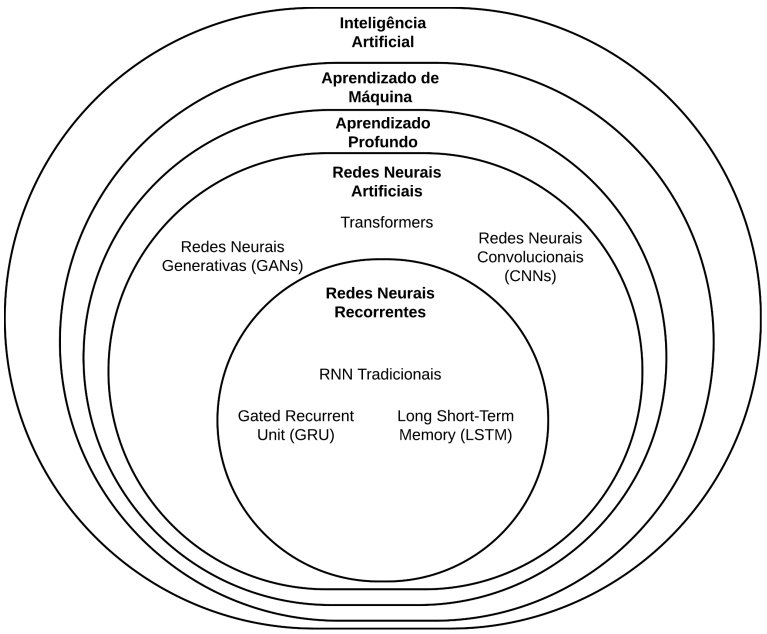
Redes Neurais Convolucionais (CNNs): Amplamente usadas em processamento de imagem e visão computacional, como em detecção de objetos, reconhecimento facial e diagnóstico médico por imagem. As CNNs são projetadas para extrair automaticamente características de imagens por meio de filtros que

capturam detalhes e padrões, desde bordas e texturas até elementos mais complexos(LeCun et al., 1998).

Redes Neurais Generativas (GANs): Aplicadas principalmente na geração de dados sintéticos, como imagens, música e até mesmo texto. Essas redes consistem em um modelo gerador e um discriminador que trabalham juntos para criar dados que simulam amostras reais. GANs são usadas em aplicações como criação de arte digital, melhoramento de imagens de baixa resolução e criação de personagens virtuais realistas (Goodfellow et al., 2014).

As redes neurais recorrentes (RNNs) são um tipo de arquitetura que compõem o um grupo das redes neurais artificiais, onde podemos encontrar outros tipos de arquiteturas para diversos propósitos. as RNNs se destacam por sua capacidade de processar dados sequenciais, sendo amplamente utilizadas em tarefas que envolvem séries temporais, como predição de eventos futuros, reconhecimento de fala e processamento de linguagem natural (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 9 - Deep Learning e as Redes neurais

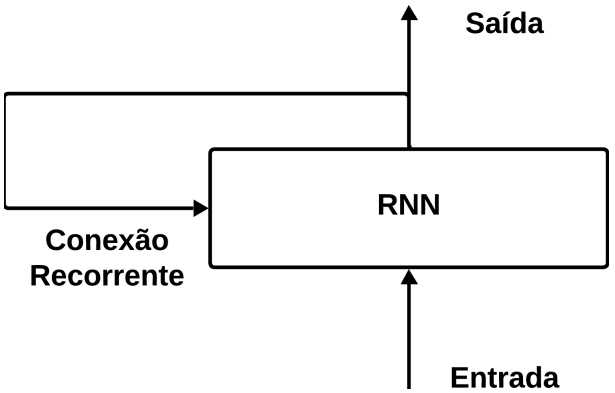


Fonte: Autor

Uma rede neural recorrente (RNN) adota o princípio de processamento sequencial e armazenamento de informações ao longo do tempo. Em uma versão mais simplificada: ela processa sequências iterando através dos elementos da

sequência e mantém um estado que contém informações relativas ao que viu até agora (CHOLLET, 2018). Com efeito, uma RNN é um tipo de rede neural que possui um loop interno (Figura 10).

Figura 10 - RNN com laço de loop



Fonte: Autor

Diferentemente das redes neurais tradicionais, que não conseguem lidar eficientemente com dados que possuem dependências temporais, as redes neurais recorrentes são projetadas para capturar essas relações de sequência nos dados. Isso é possível porque as RNNs possuem conexões internas que permitem que informações de estados anteriores sejam retidas e reutilizadas em estados subsequentes, o que é essencial para prever comportamentos ou padrões futuros (CHO et al., 2014).

As RNNs são particularmente eficazes em tarefas que envolvem séries temporais, como no caso do monitoramento de níveis de rios, onde o comportamento dos dados ao longo do tempo é crítico para fazer predições precisas. No entanto, uma das limitações das RNNs tradicionais é a dificuldade de lidar com longas dependências temporais, um problema conhecido como desvanecimento do gradiente (BENGIO et al., 1994).

Para contornar essa limitação, foram desenvolvidas variantes das RNNs, como as Long Short-Term Memory (LSTM) e as Gated Recurrent Unit (GRU). Ambas foram projetadas para lidar com o problema de dissipação do gradiente, especificamente para este problema de dependências temporais a longo prazo.

Nesta pesquisa, foi optado pela rede neural LSTM conseguem manter informações por longos períodos, sendo eficazes em modelar dependências de longo prazo, no contexto dessa pesquisa para utilizá-las para análise de longas séries temporais de dados históricos, o que as torna extremamente úteis para tarefas como predição de séries temporais (HOCHREITER; SCHMIDHUBER, 1997).

Dado o problema proposto neste estudo, a rede LSTM (Long Short-Term Memory) é adequada para a tarefa de predição dos níveis de cota do rio, considerando a natureza dos dados coletados. A LSTM pode ser eficaz para modelar séries temporais, como as variações de níveis de rios ao longo do tempo. Os detalhes sobre os dados e o processo de coleta serão descritos com mais profundidade nos capítulos subsequentes.

3.3. Descoberta de Conhecimento em Bancos de Dados (KDD)

A Descoberta de Conhecimento em Bancos de Dados ou Knowledge Discovery in Databases (KDD) é um conjunto de atividades contínuas que visa compartilhar o conhecimento extraído de bases de dados. De acordo com Fayad et al. (1996), esse processo é composto por cinco etapas principais: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; mineração de dados e interpretação e avaliação dos resultados.

Seleção de Dados: Nesta primeira etapa, são escolhidos os dados relevantes para a análise, identificando as fontes de dados e filtrando variáveis que podem ter um impacto na qualidade do modelo. A seleção inicial visa reduzir a complexidade e focar em dados que contenham as informações relevantes para o objetivo do estudo.

Pré-processamento de Dados (ou Limpeza de Dados): Aqui, os dados passam por um processo de limpeza e tratamento para corrigir problemas como dados ausentes, duplicados ou inconsistentes. Esse passo é fundamental para garantir que a análise e os modelos não sejam comprometidos por erros ou ruídos nos dados.

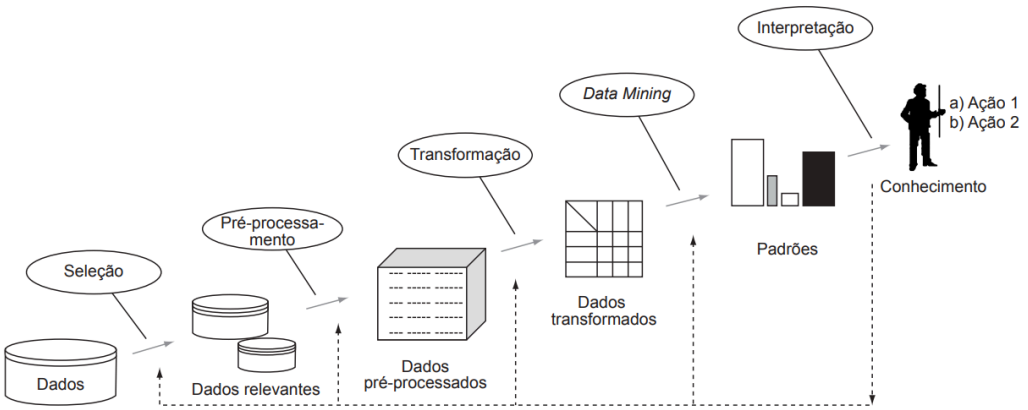
Transformação de Dados: Nesta etapa, os dados são transformados para formatos que facilitam a análise e a modelagem. Podem ser realizadas operações como normalização, agregação, construção de novos atributos, entre outros. A

transformação é essencial para que os dados estejam estruturados da forma mais adequada possível para o método de mineração que será usado.

Mineração de Dados: A etapa central do KDD, onde são aplicados algoritmos e técnicas para extrair padrões e tendências dos dados. Esta fase pode incluir a aplicação de modelos de aprendizado de máquina, como redes neurais, LSTM ou algoritmos de classificação e agrupamento, dependendo dos objetivos do projeto.

Interpretação e Avaliação: Na última etapa, os resultados obtidos na mineração de dados são interpretados e avaliados para verificar sua validade e utilidade. Esta fase envolve a análise dos padrões encontrados para identificar se eles realmente representam conhecimento valioso e contribuem para a resolução do problema inicial. É aqui que se documentam as descobertas e insights derivados do processo. A figura a seguir ilustra todo o processo KDD.

Figura 11 - Etapas do processo KDD de Fayyad et al.(1996)



Fonte:
https://www.researchgate.net/figure/Figura-1-Etapas-do-processo-KDD-Fayyad-et-al-1996_fig1_228435521

3.4. Long short-term memory (LSTM)

O Long Short-Term Memory (LSTM) foi desenvolvido por Hochreiter e Schmidhuber em 1997 como resultado de suas pesquisas sobre o problema do gradiente de fuga nas Redes Neurais Recorrentes (RNNs). Embora seja uma variação das RNNs, as LSTMs utilizam uma abordagem para capturar de maneira eficaz as dependências temporais de longo prazo em séries temporais. Além disso, pode ser combinado com mecanismos de atenção, que dão mais peso a determinados pontos importantes da série temporal (ZHANG et al., 2020).

Essa arquitetura apresenta várias vantagens, como a capacidade de focar em eventos relevantes e desconsiderar informações irrelevantes ou redundantes. A combinação pode melhorar o desempenho em tarefas onde os dados apresentam dependências temporais fortes e específicas, como predição de níveis de rios, influenciados por fenômenos sazonais ou eventos climáticos.

A aplicação da LSTM pode enfrentar algumas limitações, especialmente devido à quantidade limitada de dados históricos. Como a LSTM é projetada para capturar dependências temporais complexas, a falta de dados pode dificultar a detecção de padrões sazonais e tendências de longo prazo, essenciais em séries temporais hidroclimatológicas. Além disso, com poucos dados, o modelo tende a sofrer overfitting, ajustando-se bem aos dados de treinamento, mas falhando em generalizar para novos cenários ou variações.

Essas limitações podem impactar a capacidade de previsão do modelo, tornando-o menos robusto em condições de ruído e menos preciso em detectar ciclos sazonais. Para mitigar esses problemas, seria ideal aumentar o volume de dados, incorporar técnicas de regularização, e considerar modelos híbridos ou modelos mais simples como as GRUs ou ARIMA, que possam capturar melhor as características específicas da série temporal com dados limitados. Essas abordagens ajudam o modelo a evitar o overfitting e melhorar a generalização das previsões.

Nos parágrafos a seguir será feita uma breve explicação dessa arquitetura.

Entrada (Input)

O LSTM recebe dois tipos de entrada em cada etapa de tempo:

- Input atual (x_t)Dados da sequência atual que alimentam a célula LSTM.

- Estado oculto anterior (h_{t-1})Informação processada na etapa anterior que carrega informações relevantes acumuladas no tempo.

- Estado da célula anterior (C_{t-1})Armazena a "memória de longo prazo" da rede e transporta informações de uma etapa para outra.

Esses dois estados $(h_{t-1}$ e (C_{t-1}) são essenciais para manter as dependências temporais, garantindo que as informações relevantes ao longo da sequência possam ser retidas ou descartadas.

Forget Gate (Porta de Esquecimento)

A primeira etapa da célula LSTM é a forget gate (função de esquecimento), que decide quais informações do estado da célula anterior $((C_{t-1}$ devem ser descartadas. Esta porta é controlada por uma função sigmóide $(\sigma$ que gera valores entre 0 e 1:

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

- f_t é o vetor de saída da forget gate.
- W_f são os pesos associados à porta.
- b_f é o bias.

Se o valor de $(f_t$ for próximo de 0, as informações anteriores serão descartadas; se for próximo de 1, serão retidas. Isso é crucial para permitir que a rede "esqueça" informações desnecessárias ao longo do tempo.

Input Gate (Porta de Entrada)

A segunda etapa envolve a input gate (porta de entrada), que controla quais novas informações serão adicionadas ao estado da célula. Esta porta também é gerida por uma função sigmóide que determina quais partes da nova entrada serão atualizadas:

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

- i_t é o vetor de saída da input gate.

Além disso, uma tangente hiperbólica (tanh) é aplicada para criar um novo vetor de possíveis valores de estado da célula candidatos $\tilde{C}_t$ que podem ser adicionados ao estado atual:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

A combinação dessas duas operações permite que o LSTM determine quais informações da nova entrada serão armazenadas no estado da célula.

Atualização do Estado da Célula

Depois que as decisões de "esquecimento" e "entrada" são feitas, o estado da célula (C_t) é atualizado. O novo estado da célula é uma combinação das informações retidas $(f_t \cdot C_{t-1})$ e das novas informações adicionadas $(i_t \cdot \tilde{C}_t)$:

$$_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

Essa etapa é o coração do LSTM, pois o estado da célula (C_t) carrega as informações que serão transmitidas ao longo do tempo, agindo como uma memória de longo prazo que pode ser modificada seletivamente.

.Output Gate (Porta de Saída**)**

A output gate (porta de saída) controla quais informações do estado da célula (C_t) serão enviadas para o próximo estado oculto (h_t). Primeiro, uma função sigmóide decide quais partes do estado da célula serão emitidas:

$$\emptyset_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Em seguida, a tangente hiperbólica é aplicada ao estado da célula atualizado (C_t) para restringir os valores entre -1 e 1, e o resultado é multiplicado por (o_t) para gerar o novo estado oculto:

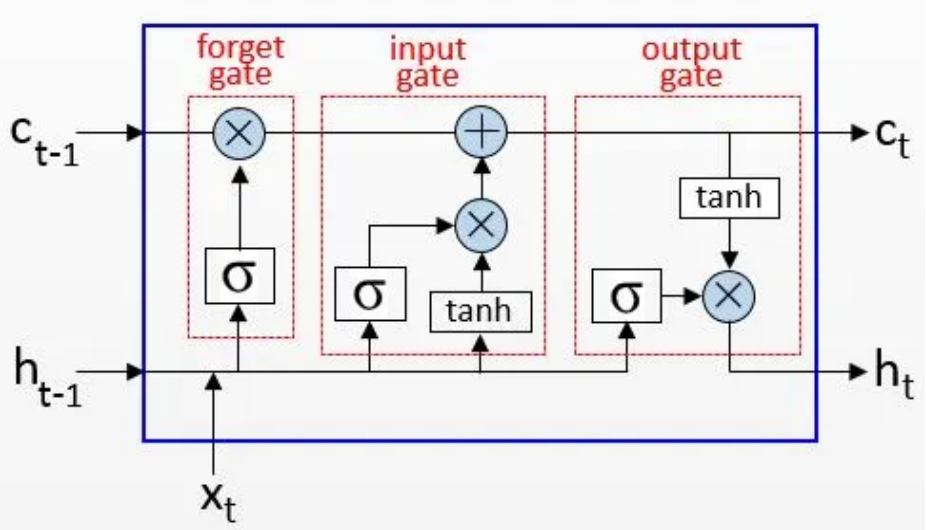
$$h_t = o_t \cdot \tanh(C_t)$$

O estado oculto (h_t) carrega as informações processadas que serão passadas para a próxima etapa de tempo, enquanto o estado da célula (C_t) armazena a "memória de longo prazo".

Em resumo:
- Forget Gate: Decide o que esquecer do estado da célula anterior.
- Input Gate: Determina quais novas informações serão armazenadas na célula.
- Output Gate: Decide quais informações do estado da célula serão usadas como saída e enviadas para a próxima etapa.

Esses três componentes trabalham em conjunto para permitir que o LSTM mantenha informações úteis ao longo do tempo, enquanto descarta aquelas que não são mais relevantes, a figura a seguir ilustra a arquitetura descrita.

Figura 12 - Diagrama do modelo LSTM



Fonte: https://dwbi1.wordpress.com/2021/08/07/recurrent-neural-network-rnn-and-lstm/

3.5. Métricas de Avaliação

3.5.1. Mean Absolute Error (MAE)

O Erro Médio Absoluto (MAE) é a média das diferenças absolutas entre os valores previstos e os valores reais. Ele mede o quanto, em média, as previsões estão distantes dos valores reais, independentemente da direção do erro.

Fórmula:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\text{Previsao}_i - \text{Valor Real}_i|$$

Onde:

- ν é o número total de amostras.
- Previsao_i é o valor previsto pelo modelo para a (i)-ésima amostra.
- Valor Real_i é o valor real para a (i)-ésima amostra.

Um MAE menor indica que o modelo está fazendo previsões mais próximas dos valores reais. É uma métrica simples que não é muito sensível a grandes erros, pois apenas considera as distâncias absolutas.

3.5.2. Mean Squared Error (MSE)

O Erro Quadrático Médio (MSE) é a média dos quadrados das diferenças entre os valores previstos e os valores reais. Essa métrica dá mais peso aos grandes erros, pois as diferenças são elevadas ao quadrado.

Fórmula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\text{Previsao}_i - \text{Valor Real}_i)^2$$

O MSE é mais sensível a grandes erros, o que pode ser útil para identificar previsões que estão significativamente fora do esperado. Um MSE menor indica um ajuste melhor do modelo, mas, devido ao seu comportamento, pode ser afetado por valores discrepantes (outliers).

3.5.3. Root Mean Squared Error (RMSE)

O Raiz do Erro Quadrático Médio (RMSE) é simplesmente a raiz quadrada do MSE. Ele também penaliza mais os grandes erros, mas é mais fácil de interpretar, pois está na mesma unidade que os valores previstos.
Fórmula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{Previsao}_i - \text{Valor Real}_i)^2}$$

O RMSE fornece uma medida direta de quanto, em média, os valores previstos diferem dos valores reais, levando em consideração a magnitude dos erros. Como é mais sensível a grandes desvios, valores menores de RMSE indicam previsões mais precisas.

4. METODOLOGIA

4.1. O objeto de estudo

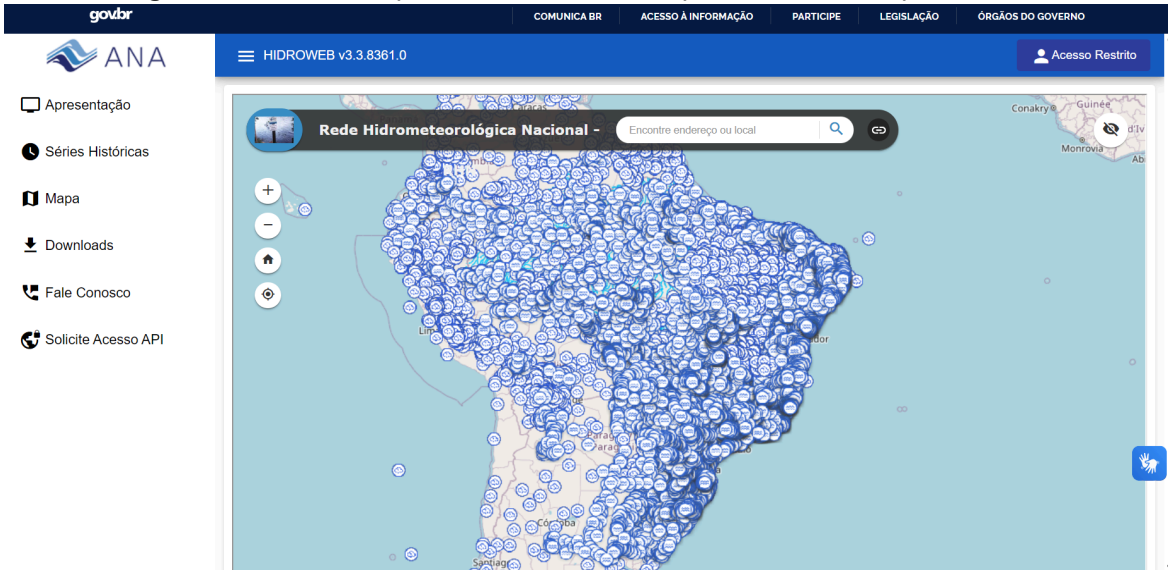
O desenvolvimento desse estudo utilizou dados de três estações em pontos distintos do rio Tapajós, a primeira localizada na cabeceira do rio, a segunda na parte central e a terceira estendendo-se até a foz. Esses pontos podem ser considerados importantes para compreender o comportamento hidrológico do rio, pois abrangem diferentes regiões que apresentam dinâmicas fluviais distintas. Essa abordagem permitirá a observação detalhada de variações sazonais e comportamentais em cada trecho, oferecendo uma visão mais abrangente sobre o impacto de eventos climáticos e mudanças sazonais nos níveis de cota ao longo do rio.

4.2. Aquisição dos dados

Os dados coletados estão disponíveis na hidroweb, uma plataforma web integrada à Rede Hidrometeorológica Nacional – RHN pertencente à Agência Nacional de Águas (ANA), é um conjunto das estações de hidrológicas mantidas por instituições públicas e privadas, voltado à geração contínua de dados representativos e confiáveis sobre os recursos hídricos nacionais.

Acessando o site, temos acesso ao mapa com todas as estações cadastradas no banco de dados da ANA. Elas são representadas por ícones azuis: o ícone de nuvem representa as estações pluviométricas (que coletam dados de chuva), enquanto o ícone de onda representa as estações fluviométricas (que coletam dados do rio) (Figura 12).

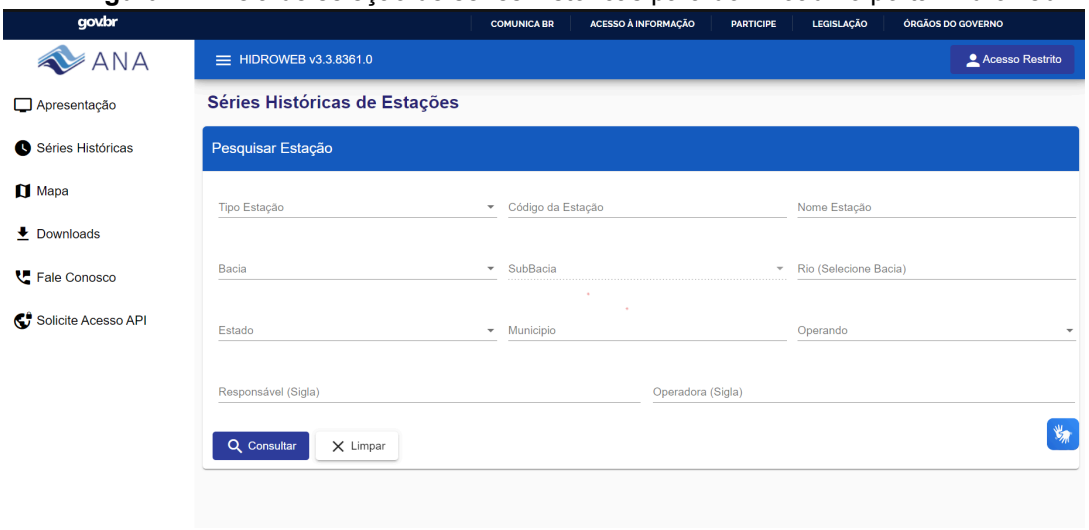
Figura 13 - Tela de mapa das séries históricas para download no portal Hidroweb



Fonte: https://www.snirh.gov.br/hidroweb/mapa

Obtendo o código de identificação da estação na aba de séries históricas (Figura 14), basta inseri-lo no campo 'Código da Estação' e escolher uma das três opções de formato: TXT, CSV ou MB. Após isso, é só baixar o arquivo, que virá em formato compactado (ZIP).

Figura 14 - Tela de seleção de séries históricas para download no portal Hidroweb



Fonte: https://www.snirh.gov.br/hidroweb/serieshistoricas

Inicialmente foi feito um levantamento de quantas estações coletam dados do rio Tapajós, e foi descoberto que há um total de 20 estações fluviométricas e 16 pluviométricas. Dentre todas as estações vistas, apenas três estações fluviométricas foram selecionadas.

As três estações de monitoramento estão localizadas em pontos distintos ao longo do rio Tapajós como supracitado, para especificar, a primeira está localizada na Barra do São Manuel, na cabeceira do rio; a segunda em Itaituba, na região central; e a terceira, em Santarém, na foz. A Figura 14 mostra a localização exata das estações no mapa, enquanto a Tabela 2 apresenta as coordenadas específicas de cada uma delas.

Figura 15 - Localização das estações fluviométricas no mapa

Fonte: Autor

Tabela 2 - Estações fluviométricas

Nº	Código	Estação	Coordenadas
1	17430000	Barra do São Manuel	-7.3397, -58.1553
2	17730000	Itaituba	-4.2756, -55.9822
3	17900000	Santarém	-2.4136, -54.7378

Fonte: Autor

O formato escolhido foi o CSV (valores separados por vírgulas). Esse formato tabular é adequado para a pesquisa, pois permite o armazenamento estruturado dos dados em tabela. Entre os arquivos CSV das estações de fluviometria disponíveis, encontram-se os dados de cotas, vazão, qualidade da água, perfil transversal, sedimentos, curva de descarga e resumo de descarga. Com os dados de cota, é possível ter uma ideia do comportamento do nível do rio ao longo de um determinado período. A tabela a seguir (Tabela 3) apresenta as informações desse arquivo de cotas, incluindo as colunas e suas descrições.

Tabela 3 - Descrição das colunas do arquivo de dados fluviométricos de cota da estação de hidro-telemetria.

Coluna	Descrição
EstacaoCodigo	Código identificador único da estação de hidrotelemetria.
NivelConsistencia	Indica o nível de consistência dos dados
Data	A data da medição.
Hora	A hora da medição (pode estar ausente em algumas linhas)
MediaDiaria	Média diária das vazões
TipoMedicaoCotas	O tipo de medição das vazões
Maxima	O valor máximo da vazão registrado para o dia
Minima	O valor mínimo da vazão registrado para o dia.
Media	A média dos valores registrados da vazão para o dia.
DiaMaxima	O dia em que foi registrado o valor máximo da vazão.
DiaMinima	O dia em que o valor mínimo de cota foi registrado.
MaximaStatus	Um indicador do status ou qualidade do valor máximo de cota.
MinimaStatus	O status ou qualidade do valor mínimo de cota.
MediaStatus	Indicador do status da média geral dos valores de cota.
MediaAnual	A média anual dos valores de cota.
MediaAnualStatus	O status ou qualidade da média anual de cotas.
Cota01 a Cota31	Valores de cota para cada dia do mês, mostrando as medições diárias do nível de água.

Fonte: Autor

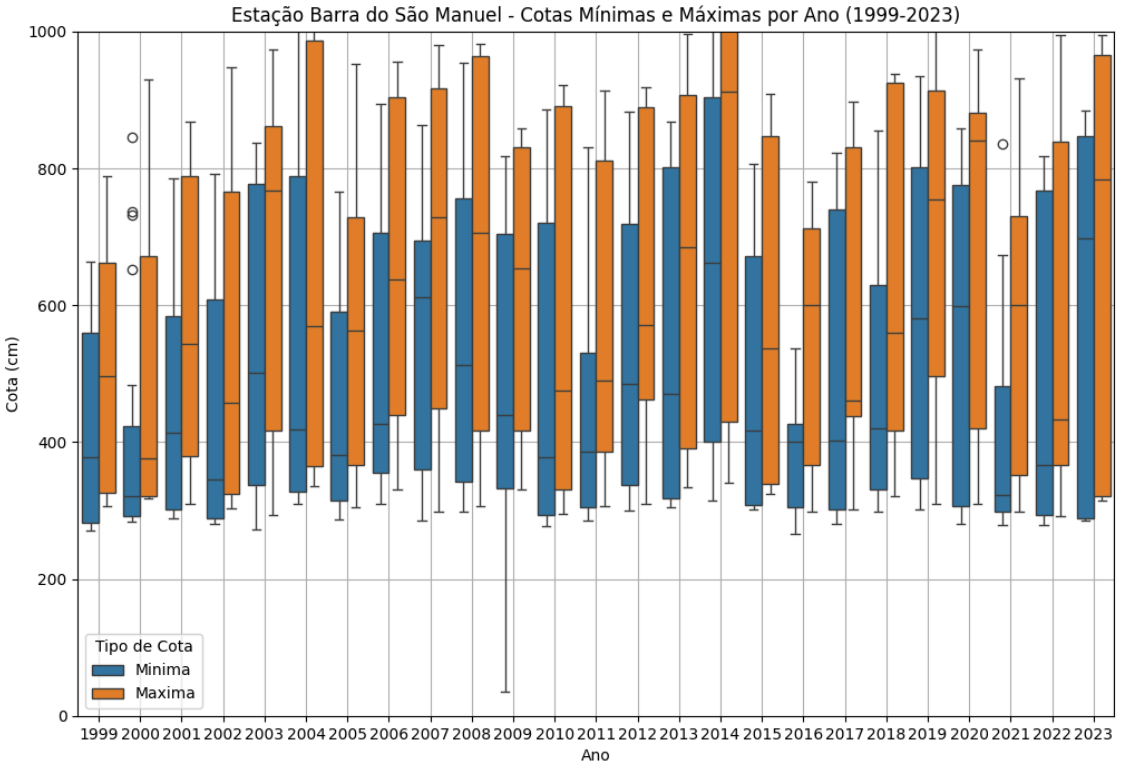
4.3. Ambiente de desenvolvimento

Todo o processamento dos dados, bem como a construção e execução do modelo LSTM, foi realizado utilizando a plataforma Google Colab, uma ferramenta de processamento em nuvem que permite a execução de códigos em arquivos no formato ipynb, é um tipo de arquivo usado pelo Jupyter Notebook, uma aplicação interativa de código aberto que permite criar e compartilhar documentos que contêm código executável, visualizações e explicações textuais. O .ipynb é amplamente utilizado para análise de dados, aprendizado de máquina e outros tipos de computação científica.

As etapas foram implementadas por meio de scripts em Python, com o apoio de bibliotecas de machine learning, como Scikit-learn, Keras e TensorFlow, garantindo a eficiência e a precisão nas análises e predições realizadas.

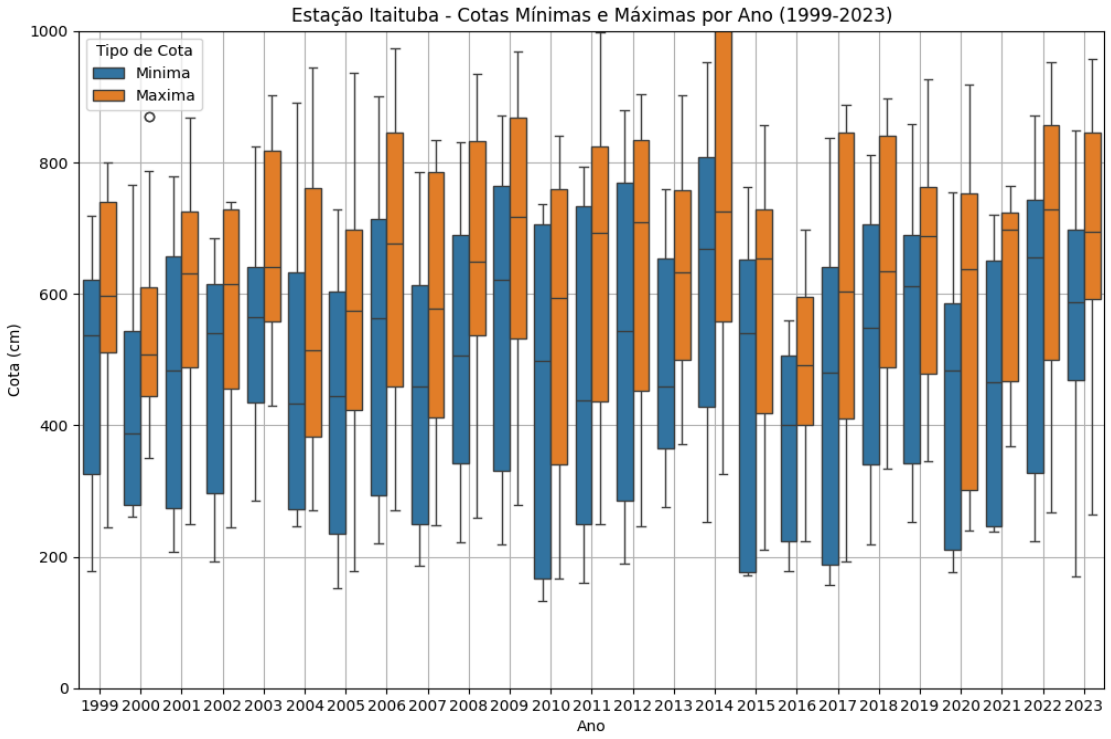
Uma visão geral acerca dos dados brutos foram realizadas e serão mostradas as imagens a seguir (Figuras 15, 16 e 17), cada gráfico representa uma média de cotas mínimas e máximas durante todo o período de 1999 a 2023.

Figura 16 - Boxplot das mínimas e máximas da estação Barra do São Manuel



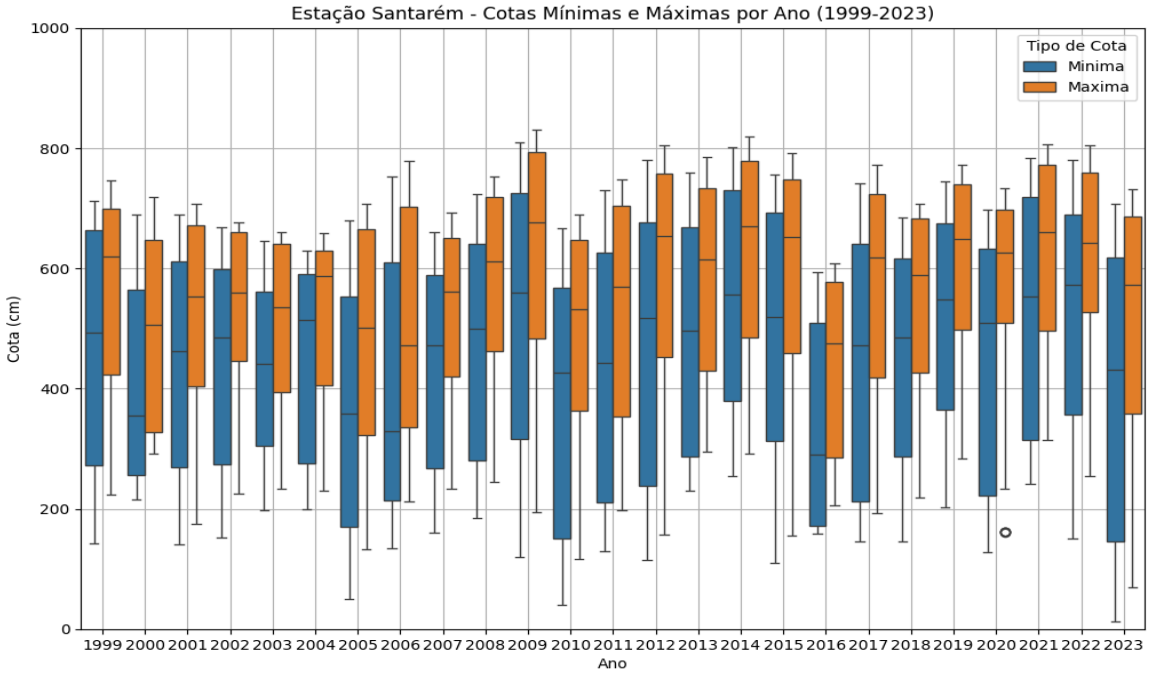
Fonte: Autor

Figura 17 - Boxplot das mínimas e máximas da estação Itaituba



Fonte: Autor

Figura 18 - Boxplot das mínimas e máximas da estação Santarém



Fonte: Autor

Observa-se que, entre as três estações, a Barra do São Manuel se mantém em níveis mais altos tanto nas cotas máximas quanto nas mínimas em relação às

outras duas. Sendo a nascente do rio, com dois tributários, os níveis de cota dessa estação encontram-se dentro dos valores esperados.

A estação de Santarém apresenta níveis menores tanto para mínimas quanto para máximas; sendo o ponto de vazão do rio, é esperado que tenha os menores valores entre as três. A estação de Itaituba mantém-se em uma média de cotas mínimas e máximas, oscilando entre valores altos, como ocorre na estação de São Manuel, e valores baixos, como na estação de Santarém — um comportamento típico de um trecho médio do rio.

4.4. Aplicação do método KDD

As etapas a seguir descrevem o passo a passo do processamento dos dados.

- Seleção

Como supracitado o levantamento de estações foi necessário realizar a seleção das três estações com os dados utilizados.

- Pré processamento

A seleção dos atributos foi realizada com o objetivo de focar nas variáveis mais relevantes para a análise preditiva dos níveis do rio. As colunas de data e cotas dos dias 01 a 31 foram escolhidas porque a data permite a análise temporal dos dados, essencial para capturar padrões sazonais ou cíclicos, enquanto as cotas fornecem as informações cruciais sobre o comportamento do nível do rio ao longo do tempo.

A restrição aos dias de 01 a 31 visa garantir a consistência e a comparabilidade dos dados dentro de um intervalo específico, permitindo que o modelo capture as flutuações diárias de maneira mais eficiente, sem interferência de dados desnecessários ou irrelevantes para a previsão.

O período de 1999 a 2023 foi selecionado para assegurar a representatividade de uma ampla série temporal, permitindo que o modelo aprenda padrões de longo prazo.Esse intervalo temporal também é suficiente para cobrir diferentes condições hidrológicas e climáticas, garantindo a robustez e a generalização do modelo.

- Transformação

A normalização é um processo que compõe a etapa da transformação de dados, o objetivo é tornar os dados mais apropriados para aplicação de algoritmos de mineração de dados e de redes neurais. Ela evita a saturação dos neurônios para uma rede de múltiplas camadas como o LSTM.

As redes LSTM geralmente se beneficiam da normalização Min-Max, pois as funções de ativação (como tanh e sigmóide) operam melhor em valores entre 0 e 1. Isso pode melhorar o tempo de convergência do modelo e o desempenho final, sendo assim, neste estudo foi utilizada a normalização Min-Max, para fins de conveniência, o modelo LSTM converge melhor com a escala de 0 a 1.

A normalização é particularmente importante em redes neurais porque os pesos iniciais dos neurônios são geralmente pequenos, e dados com escalas muito diferentes podem dificultar a convergência do modelo.

Fórmula de normalização Min-Max:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

- x_{norm}: o valor normalizado.
- x: o valor original.
- $x_{\min}$: o valor mínimo da coluna.
- $x_{\max}$: o valor máximo da coluna.

Todos os valores originais x são ajustados para uma faixa definida, geralmente entre 0 e 1. O valor mais baixo da série se torna 0 e o valor mais alto se torna 1. Todos os valores intermediários são escalados proporcionalmente dentro dessa faixa. Para isso é executado o comando *scaler = MinMaxScaler()* sobre o conjunto de dados que a normalização será feita.

Ao final do processamento, executando o comando *df.head()* obtemos o retorno dos nossos dados normalizados, as imagens a seguir (Figura 18 e 19) mostram os dados brutos e pós normalizados.

Figura 19 - Amostra dos dados brutos

	Data	Cota01	Cota02	Cota03	Cota04	Cota05	Cota06	Cota07	Cota08	Cota09
4	2023-01-08	597.0	592.0	587.0	585.0	579.0	575.0	569.0	565.0	563.0
3	2023-01-09	447.0	443.0	439.0	438.0	436.0	428.0	419.0	413.0	397.0
2	2023-01-10	166.0	161.0	156.0	148.0	137.0	125.0	107.0	94.0	83.0
1	2023-01-11	42.0	40.0	42.0	47.0	46.0	39.0	24.0	13.0	12.0
0	2023-01-12	72.0	75.0	74.0	77.0	71.0	63.0	62.0	62.0	73.0

Fonte: Autor

Figura 20 - Amostra dos dados pós normalizados

	Data	Cota01	Cota02	Cota03	Cota04	Cota05	Cota06	Cota07	Cota08	Cota09	...
0	2023-12-01	0.038168	0.044643	0.041026	0.038810	0.032383	0.030888	0.047919	0.060870	0.075495	...
1	2023-11-01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
2	2023-10-01	0.157761	0.154337	0.146154	0.130660	0.117876	0.110682	0.104666	0.100621	0.087871	...
3	2023-09-01	0.515267	0.514031	0.508974	0.505821	0.505181	0.500644	0.498108	0.496894	0.476485	...
4	2023-08-01	0.706107	0.704082	0.698718	0.695990	0.690415	0.689833	0.687264	0.685714	0.681931	...

Fonte: Autor

Valor Normalizado de 0: Representa o nível do rio correspondente ao valor mínimo observado nos seus dados históricos (entre 1999 e 2023). Ou seja, quando o valor está próximo de 0, o nível do rio está em uma das suas cotas mais baixas.

Valor Normalizado de 1: Representa o nível máximo observado nos seus dados históricos. Quando o valor está próximo de 1, significa que o nível do rio está em uma das suas cotas mais altas.

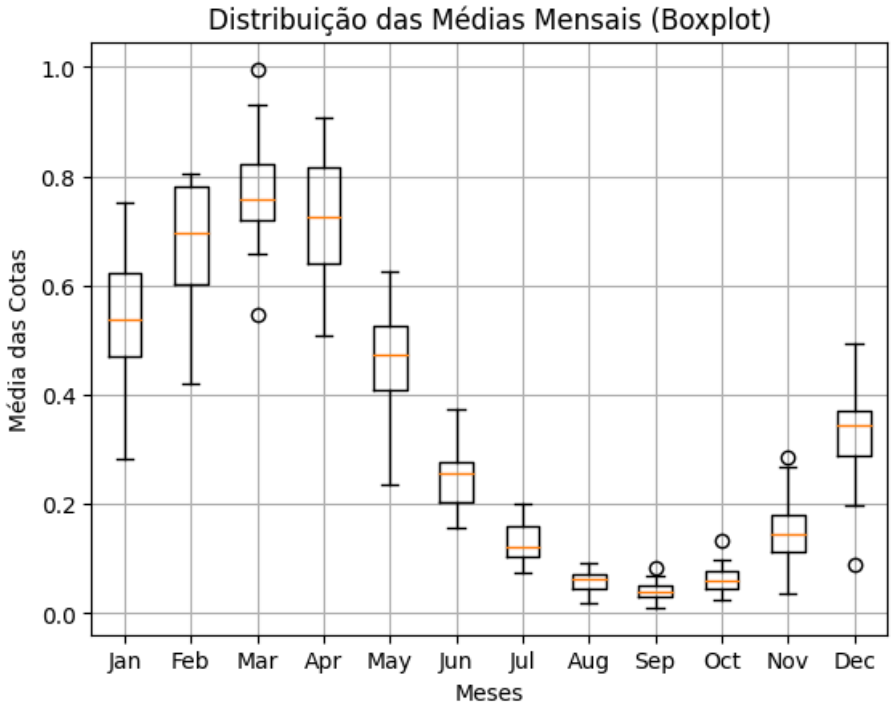
Valores entre 0 e 1: Esses valores indicam uma posição relativa do nível do rio em comparação com os extremos do período observado. Por exemplo, um valor normalizado de 0.5 indicaria que o nível do rio está em uma posição intermediária entre o mínimo e o máximo observado nos dados.

- Mineração e Avaliação dos dados

O objetivo desta etapa é analisar as minúcias e as tendências do comportamento dos dados de cotas pós-processados. Essa análise é crucial para compreender a amplitude e a distribuição dos dados, o que ajudará a ajustar e configurar corretamente a rede neural LSTM para receber esses dados como entrada. A avaliação das variáveis, como as flutuações diárias e sazonais nas cotas, possibilita a identificação de padrões temporais e a verificação da consistência dos dados. Essas observações são fundamentais para definir os parâmetros de treinamento da rede neural, garantindo que ela seja capaz de aprender de forma eficaz a dinâmica dos níveis do rio e, assim, realizar previsões precisas. As imagens a seguir (Figura 19 a 24) mostram as análises desses dados transformados.

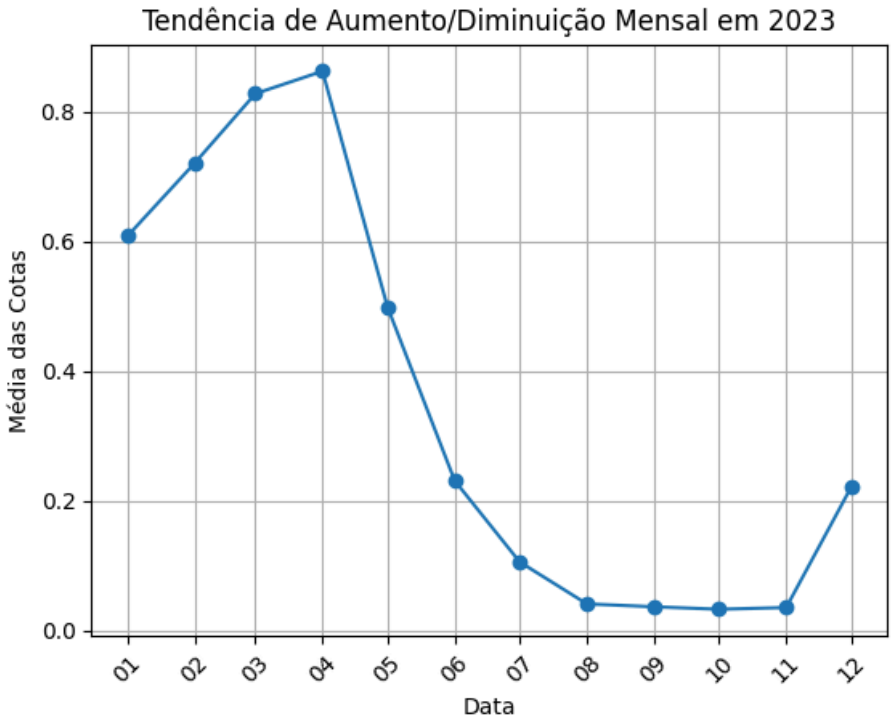
Estação Barra do São Manuel

Figura 21 - Estação Barra do São Manuel: Médias de cotas mensais de 1999 a 2023



Fonte: Autor

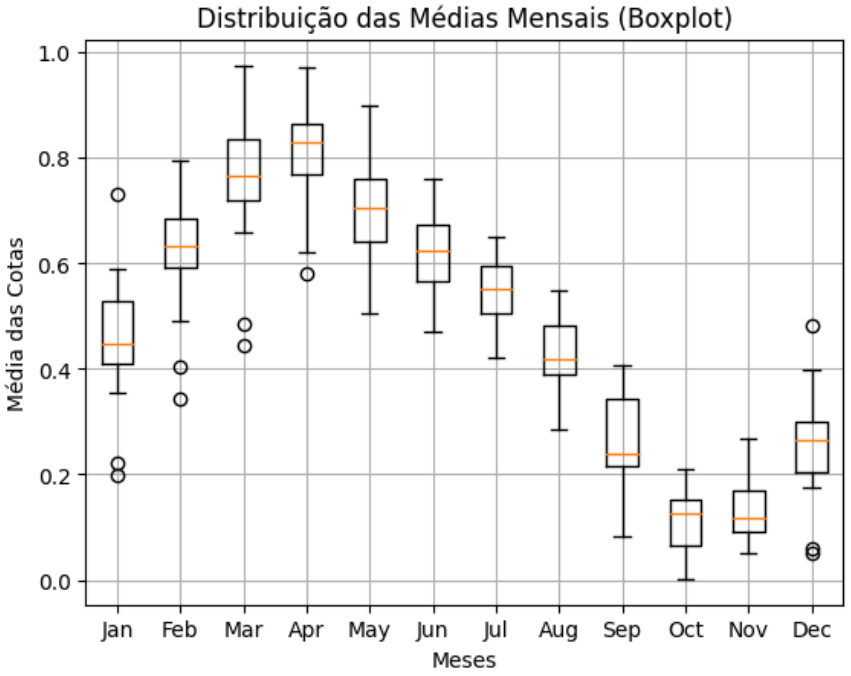
Figura 22 - Estação Barra do São Manoel: Tendências mensais de 1999 a 2023



Fonte: Autor

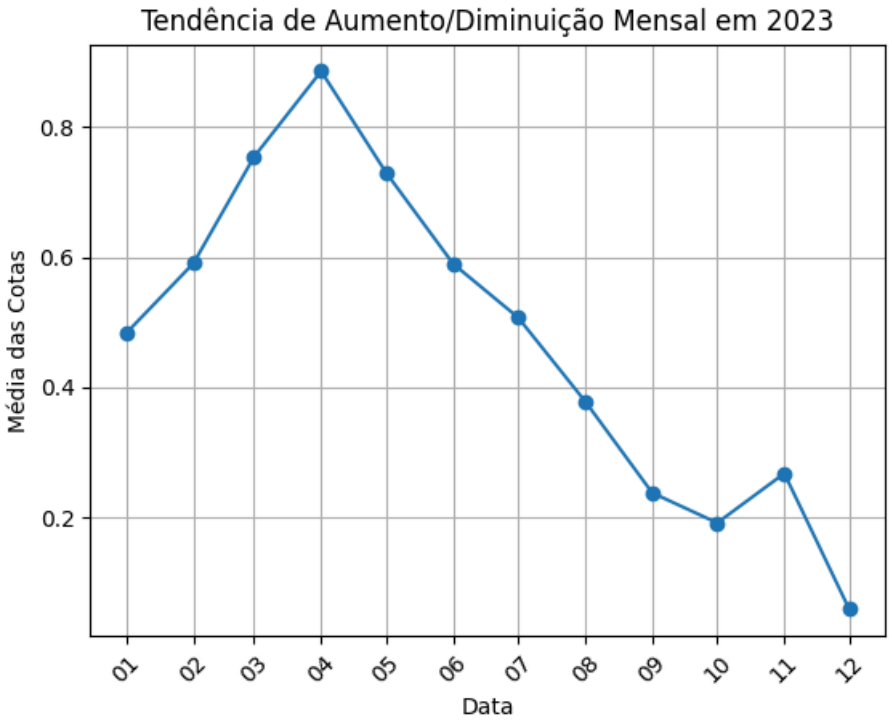
Estação Itaituba

Figura 23 - Estação Itaituba: Médias de cotas mensais de 1999 a 2023



Fonte: Autor

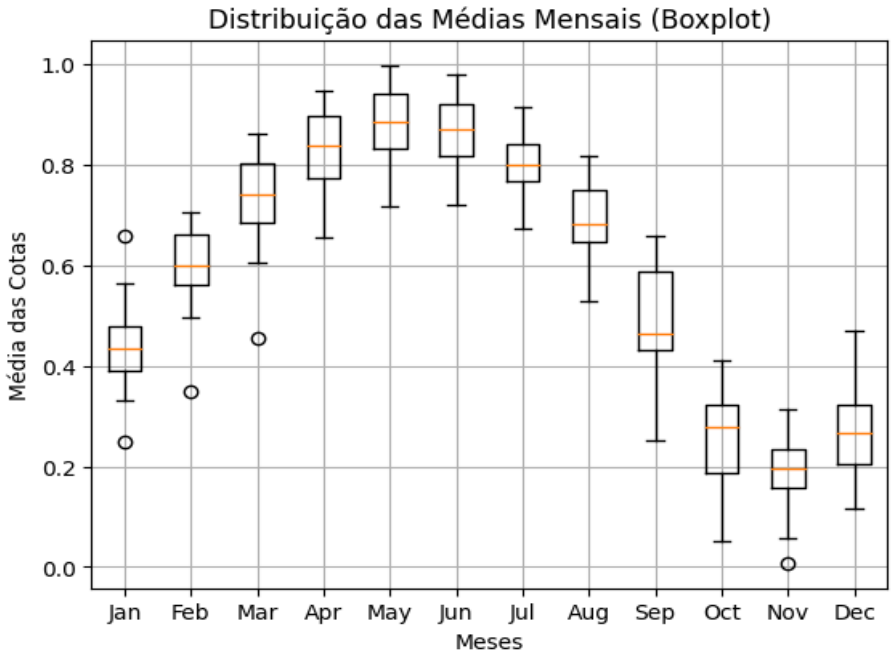
Figura 24 - Estação Itaituba: Tendências mensais de 1999 a 2023



Fonte: Autor

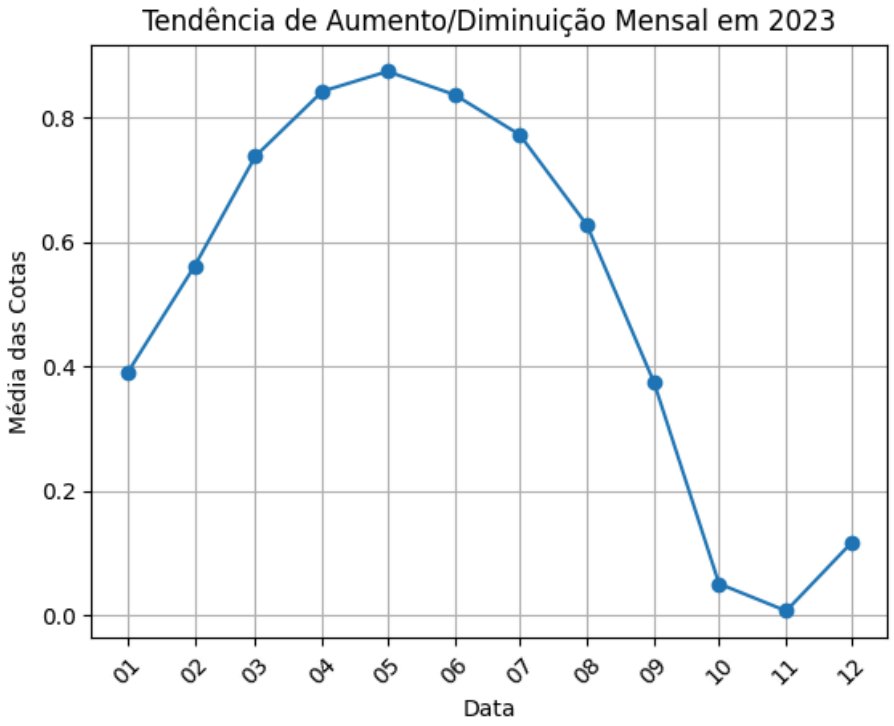
Estação Santarém

Figura 25 - Estação Santarém: Médias de cotas mensais de 1999 a 2023



Fonte: Autor

Figura 26 - Estação Santarém: Tendências mensais de 1999 a 2023



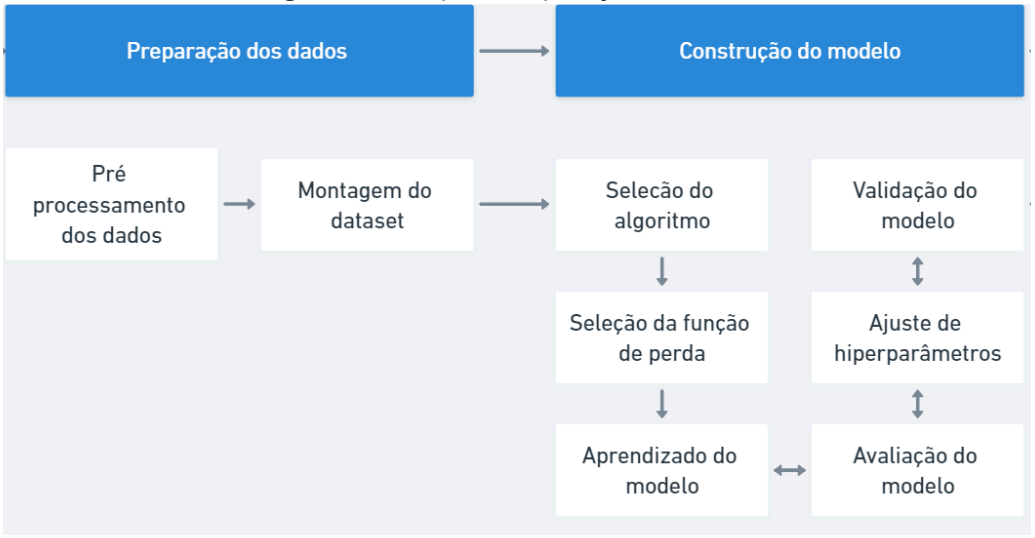
Fonte: Autor

Observa-se padrões de comportamento nos dados uma tendência clara de aumento das cotas a partir de janeiro, atingindo um pico entre abril e julho, e uma queda significativa a partir de agosto até alcançar valores mínimos em outubro e novembro. Essa tendência sazonal ocorre devido ao período de baixa precipitação e fatores externos

4.5. Aplicação do modelo LSTM

As etapas da preparação, construção, treinamento e avaliação do do modelo estão ilustradas na figura 25 e nos parágrafos a seguir.

Figura 27 - Etapas da aplicação do modelo LSTM



Fonte: Adaptado de
WOLF, Andrew. The Machine Learning Simplified: A Gentle Introduction to Supervised Learning. Edição em inglês. [eBook Kindle]. 2022.

Uma etapa de preparação dos dados é precedida a construção do modelo, e é necessária para formatar os dados para o treinamento do modelo de rede neural.

- Pré processamento dos dados

É necessário a reformatação dos dados em sequências temporais, para que se adequem as entradas *inputs* da LSTM, transformando os dados em uma forma que o modelo consiga entender. Isso envolve a criação de sequências de entrada (timesteps) e a definição do valor que queremos prever.

Janelas de tempo (lookback) em redes neurais artificiais, define o número de passos temporais anteriores que o modelo vai usar para criar as sequências de entrada. O modelo visualiza uma janela de dados de tamanho lookback antes de prever o próximo valor. A configuração do lookback tem um papel crucial na capacidade da rede de capturar dependências temporais e tendências em dados de séries temporais, se o período de lookback for muito pequeno, o modelo pode não ser capaz de capturar padrões de longo prazo. Por outro lado, se o lookback for muito grande, a rede pode ter dificuldade em aprender as relações importantes sem sobrecarregar a memória e o poder computacional.

O valor padrão do lookback que foi utilizado é 1, isso significa que cada valor de saída depende de apenas um valor de entrada anterior. A função percorre o Data Frame da linha 0 até a linha (len(data) - lookback). Isso significa que, para cada

sequência, ele pega uma "janela" de dados com lookback linhas consecutivas, cada entrada é uma linha e a saída é a linha seguinte. Definir o lookback como 1 mantém o modelo simples, e ele só precisará aprender a relação entre um ponto de dados anterior e o próximo, reduzindo a complexidade.

- Montagem do dataset

É feita uma divisão de dados: Separá-los em conjuntos de treinamento, teste, validação (70% para treinamento, 15% teste para e 15% para validação) (Figura 26).

Figura 28 - Etapas da aplicação do modelo LSTM

```
Formato das entradas de Treinamento (X_train): (171, 1, 31)
Formato das saídas de Treinamento (y_train): (171, 31)
Formato das entradas de Validação (X_val): (35, 1, 31)
Formato das saídas de Validação (y_val): (35, 31)
Formato das entradas de Teste (X_test): (37, 1, 31)
Formato das saídas de Teste (y_test): (37, 31)
```

Fonte: Autor

Essa divisão temporal foi feita respeitando a ordem temporal dos dados , não foi realizada divisão aleatória, começam por ordem cronológica, o conjunto de treinamento se inicia a partir de 1999 e o conjunto de validação no ano de 2023.

- Construção do Modelo LSTM

Definir arquitetura a ser utilizada:
Camada LSTM (com unidades LSTM).
Camada densa (fully connected) para a saída.
Função de ativação apropriada.
Loss Function, uma função de perda adequada, como Mean Squared Error (MSE).
Otimizador como Adam.

- Descrição Estrutura do Modelo

Modelo Sequencial: O modelo é criado usando o comando models.Sequential(), que indica que as camadas serão adicionadas uma após a outra, formando uma pilha linear de camadas.

Primeira Camada LSTM:

- Tipo: LSTM (Long Short-Term Memory)
- Unidades: 64 células de memória (neurônios).
- Ativação: Função de ativação tanh, que ajuda a regular os valores entre 0 e 1.
- Input Shape: O formato de entrada esperado é (X_train.shape[1], X_train.shape[2]), onde:
    - X_train.shape[1] representa o número de timesteps (sequências temporais).
    - X_train.shape[2] representa o número de características (features) por timestep.
- Return Sequences: Configurado como True, o que significa que essa camada retorna toda a sequência de saída para ser usada pela próxima camada LSTM.

Segunda Camada LSTM:

- Tipo: LSTM
- Unidades: 32 células de memória.
- Ativação: tanh.
- Return Sequences: O padrão é False, então essa camada retorna apenas a última saída da sequência.
- Camada de Saída Densa:

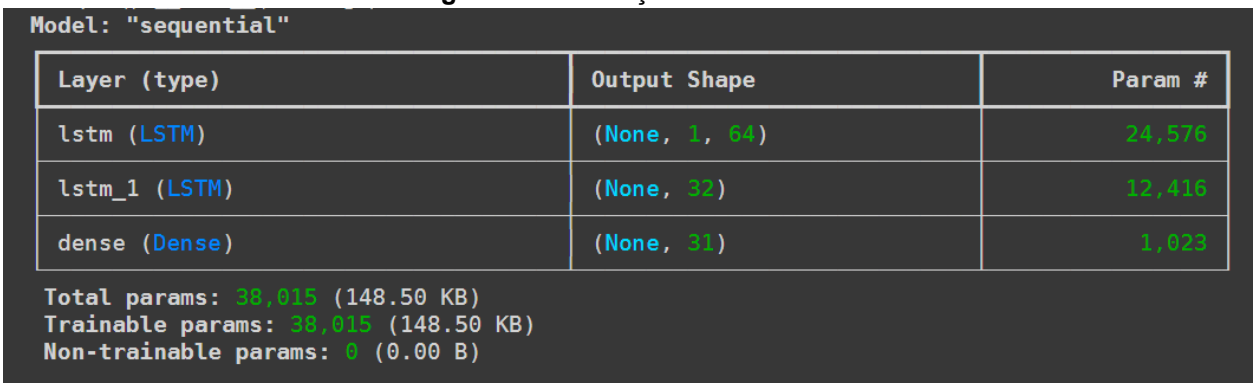
Camada totalmente conectada (Dense):

- Unidades: O número de neurônios na camada de saída é igual a y_train.shape[1], o que significa que a quantidade de saídas depende do número de variáveis-alvo que o modelo precisa prever.
- Esta camada é responsável por gerar as previsões finais.
- Compilação do Modelo
- Otimizador Adam: Frequentemente utilizado em conjuntos de dados relativamente menores, tende a convergir mais rapidamente do que outros otimizadores, ajuda a mitigar isso usando o cálculo de momento, o que estabiliza o processo de atualização dos pesos e evita flutuações bruscas.
- Taxa de aprendizado (learning_rate) de 0.001, que é uma escolha comum e eficiente para a otimização de redes neurais.

- Função de Custo: mean_squared_error (erro quadrático médio), ideal para problemas de regressão onde queremos minimizar a diferença quadrática entre os valores previstos e os reais.

Ao executar o comando *model.summary()* Obtemos a estrutura do modelo (Figura 27).

Figura 29 - Descrição do modelo LSTM



Model: "sequential"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 1, 64)	24,576
lstm_1 (LSTM)	(None, 32)	12,416
dense (Dense)	(None, 31)	1,023

Total params: 38,015 (148.50 KB)
Trainable params: 38,015 (148.50 KB)
Non-trainable params: 0 (0.00 B)

Fonte: Autor

Descrição detalhada do modelo:

Sequential: É uma classe da biblioteca Keras para criação de modelos sequenciais.

Primeira camada LSTM:

- Output Shape: (None, 1, 64)
- Número de parâmetros: 24.576
- Esta camada tem 64 unidades e está configurada para retornar uma sequência completa (return_sequences=True). A forma de saída indica que ela processa a entrada em segmentos e mantém a dimensão temporal intacta.

Segunda camada LSTM:

- Output Shape: (None, 32)
- Número de parâmetros: 12.416

- Esta camada LSTM reduz a sequência para uma representação de 32 unidades e não retorna uma sequência, apenas a última saída da série temporal.

Camada Densa (Dense):

- Output Shape: (None, 31)
- Número de parâmetros: 1.023
- Esta camada é uma camada totalmente conectada (fully connected) que gera a saída final do modelo com 31 unidades, correspondendo à quantidade de previsões que o modelo precisa fazer.

Análise dos Parâmetros

- Total de parâmetros: 38.015
- Parâmetros treináveis: 38.015
- Parâmetros não treináveis: 0

Todos os parâmetros são treináveis, o que é esperado para um modelo que não possui camadas congeladas ou pré-treinadas.

Treinamento do Modelo

- Treinamento: Alimente o modelo com os dados de treinamento e ajuste os parâmetros. Defina o número de épocas e o tamanho do batch.
- Monitoramento: Monitore a perda durante o treinamento para evitar o overfitting.

Avaliação do Modelo

As métricas MAE, MSE e RMSE são usadas para avaliar a precisão de modelos de regressão, medindo a diferença entre os valores previstos pelo modelo e os valores reais. A seguir uma descrição de cada uma.

Essas métricas são usadas em conjunto para fornecer uma visão completa sobre a precisão do modelo, ajudando a identificar não apenas a magnitude média dos erros, mas também a importância dos erros mais significativos.

Os hiperparâmetros de treinamento estão especificados na figura a seguir.

Figura 30 - Hiperparâmetros do modelo LSTM

```
history = model.fit(X_train, y_train,
                    validation_data=(X_val, y_val),
                    epochs=50,
                    batch_size=16,
                    verbose=1)
```

Fonte: Autor

Época de Treinamento: O modelo foi treinado por 50 épocas, o que significa que o conjunto de dados foi passado 50 vezes pelo modelo durante o processo de aprendizado.

Batch Size: O número de amostras que serão processadas antes de atualizar os pesos do modelo foi 16. Esse valor significa que o modelo irá processar 16 amostras antes de realizar uma atualização dos pesos.

Verbose: O valor 1 significa que informações sobre o progresso do treinamento serão exibidas em cada época. Isso pode ser útil para monitorar o treinamento em tempo real.

Considerações sobre as configurações do modelo LSTM escolhidas e os processos de ajuste:

Foi importante levar em consideração aspectos como simplicidade do modelo, a natureza dos dados, os objetivos do estudo e os resultados, as escolhas serão descritas a seguir por tópicos.

Função de Ativação: A função de ativação padrão como tanh ou sigmoid, são amplamente utilizadas em LSTMs devido à sua habilidade de comprimir os valores entre -1 e 1 (ou 0 e 1, no caso da sigmoid), o que é adequado para a tarefa de predição séries temporais, onde você quer garantir que os valores não saiam de uma faixa controlada. Funções como ReLU ou Leaky ReLU poderiam ser mais apropriadas em tarefas de classificação ou reconhecimento de imagens, mas para séries temporais com valores normalizados, funções suaves como tanh são mais apropriadas.

Número de Camadas LSTM: A arquitetura multicamada permite que o estado de uma camada seja refinado em camadas posteriores, o que é importante em previsões de séries temporais onde as dependências temporais são críticas para a

precisão. Isso é particularmente relevante para a modelagem de variações sazonais e outras flutuações nos níveis do rio ao longo do tempo.

Testes Iterativos: Durante o desenvolvimento, foram realizadas várias iterações com diferentes números de camadas e neurônios, visando maximizar a capacidade do modelo de capturar as dinâmicas dos dados de séries temporais. Camadas adicionais foram testadas para verificar se o aumento da complexidade resultava em melhorias de precisão nas previsões.

A arquitetura multicamada escolhida permite que o estado de cada camada seja refinado nas camadas subsequentes. Isso é fundamental em séries temporais, pois a informação propagada através das camadas refina as dependências temporais de curto e longo prazo

Critérios de Seleção: O critério principal para a escolha foi o desempenho nos conjuntos de validação e teste, medido pela métrica de erro quadrático médio (MSE). Configurações mais simples (menos camadas) não capturaram adequadamente as variações temporais e sazonais dos níveis do rio, enquanto arquiteturas muito complexas começaram a sofrer com o overfitting.

Normalização Min-Max: Essa normalização é particularmente útil em redes neurais como LSTM, pois as funções de ativação (como tanh e sigmoid) são sensíveis à escala dos dados. Manter os valores entre 0 e 1 (ou -1 e 1 no caso de tanh) garante que o modelo consiga aprender mais eficientemente e evitar problemas de gradiente explodido ou gradiente desaparecido.

A padronização z-score normalization poderia ser uma opção já que ajusta os dados para uma média de 0 e um desvio padrão de 1, pode ser mais apropriada em casos onde há uma distribuição normal subjacente nos dados. No entanto, os níveis do rio variam sazonalmente e podem não seguir essa distribuição, tornando a normalização Min-Max mais adequada.

Parâmetros de Treinamento: A escolha de uma taxa de aprendizado moderada (learning_rate=0.001) garante que o modelo faça atualizações consistentes evitando oscilações no erro de treinamento. Uma taxa muito alta poderia resultar em falta de convergência, enquanto uma taxa muito baixa prolonga o tempo de treinamento sem garantias de melhorias significativas.

O tamanho do lote (batch size=16) menor facilita a generalização e evita o overfitting, ao passo que batches maiores podem reduzir o ruído no gradiente, mas também podem fazer com que o modelo se ajuste muito aos padrões dos dados de

treinamento, sem generalizar bem para novos dados (como os de teste ou validação).

O objetivo principal é prever níveis do Rio Tapajós com base em dados hidroclimatológicos, e aumentar a dimensionalidade ou a quantidade de camadas densas poderia sobrecarregar o modelo com variáveis desnecessárias, desviando o foco dos padrões temporais sazonais que o LSTM foi projetado para capturar. Além disso, adicionar muitas camadas densas ou features irrelevantes poderia levar a *overfitting* e comprometer a interpretabilidade dos resultados.

As decisões de configuração do modelo LSTM, como função de ativação, normalização, número de camadas e regularização, foram orientadas pela natureza do problema de predição de séries temporais hidroclimatológicas, a quantidade e qualidade dos dados disponíveis, e o objetivo de manter um equilíbrio entre simplicidade do modelo e performance. Essas escolhas evitam que o modelo se torne excessivamente complexo, difícil de treinar ou sensível a flutuações nos dados, garantindo uma generalização robusta para as previsões futuras.

5. RESULTADOS

Com as configurações do modelo explicadas na seção anterior, foi realizado o treinamento da rede neural.

A tabela a seguir mostra os resultados das métricas de avaliação do modelo para cada conjunto de dados, de acordo com cada métrica explicadas anteriomente e o número de amostras das sequências dos dados.

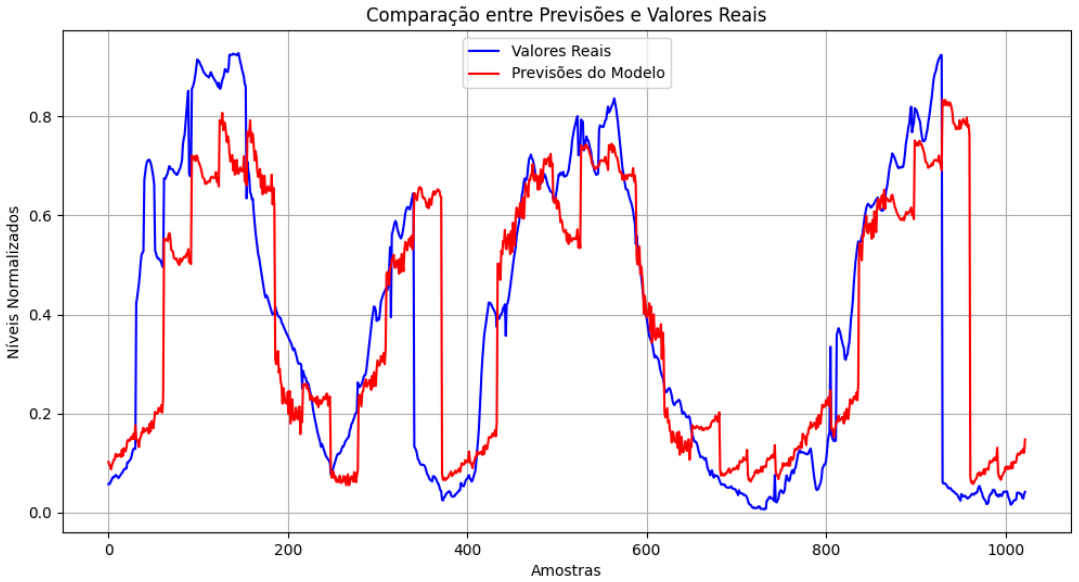
Tabela 4 - Valores das métricas de avaliação

Estação	Nº Amostras	MAE	MSE	RMSE
Barra do São Manuel	1023	0.1318	0.0429	0.2072
Itaituba	1240	0.0807	0.0225	0.1501
Santarém	1333	0.0355	0.0021	0.0460

Fonte: Autor

Estação Barra do São Manuel

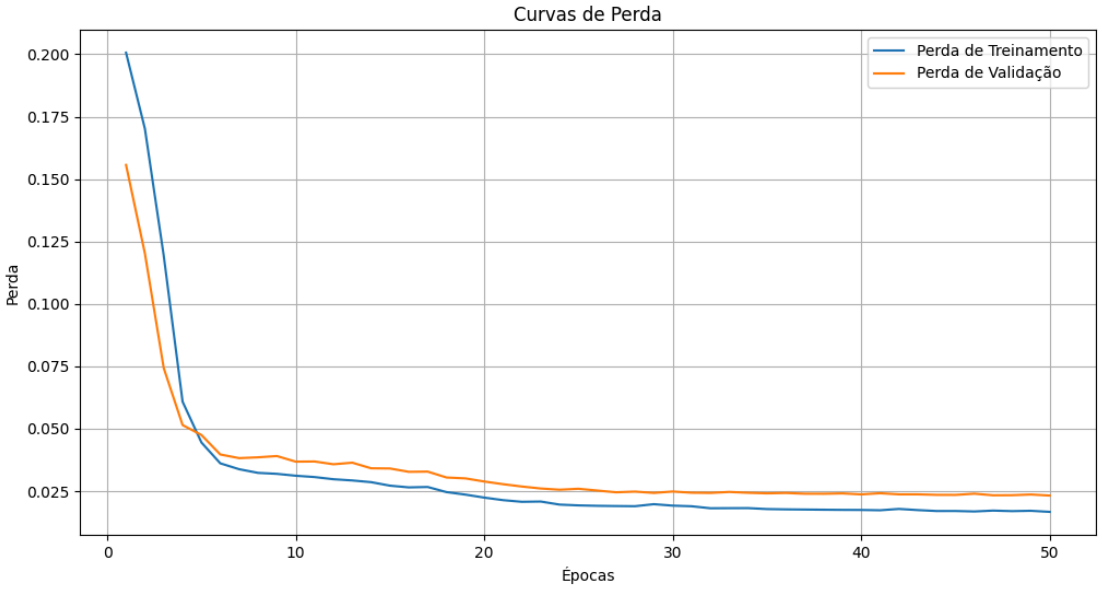
Figura 31 - Estação Barra do São Manuel: Comparação de previsões com valores reais



Fonte: Autor

A análise gráfica indica que as predições do modelo conseguem acompanhar a tendência geral dos dados, porém, há pouca convergência dos dados principalmente no começo entre as amostras de 0 a 400.

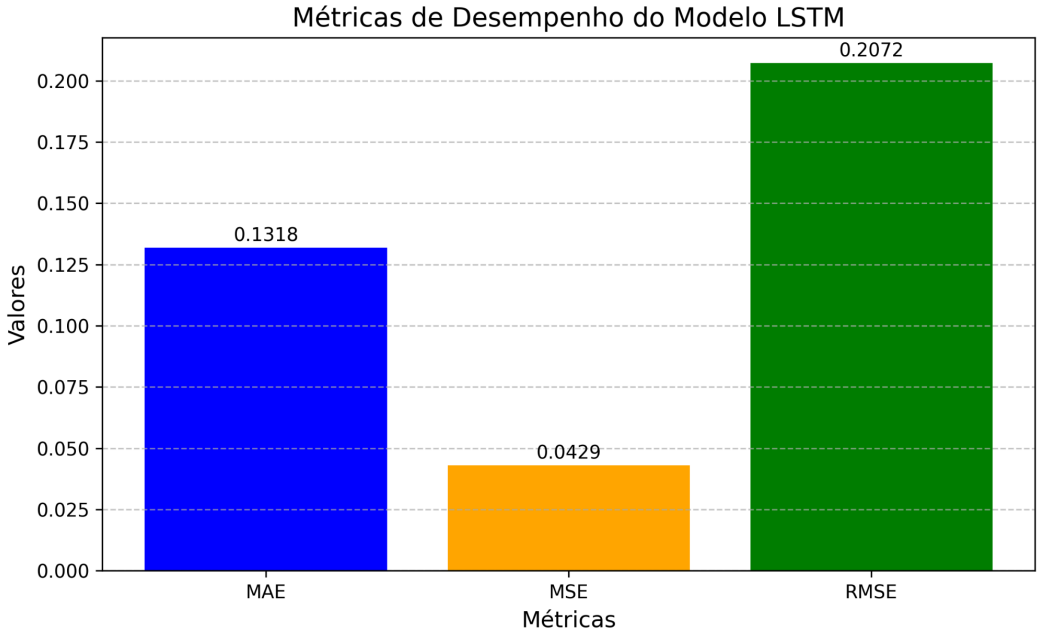
Figura 32 - Estação Barra do São Manuel: Função de perda no conjunto teste



Fonte: Autor

As curvas de perda apresentadas demonstram a convergência de um modelo durante o treinamento. A função de perda, calculada para os conjuntos de treinamento e validação, diminui significativamente nas primeiras épocas e se estabiliza em seguida, porém não há uma convergência em nenhuma época.

Figura 33 - Estação Barra do São Manuel: Resultados das métricas de desempenho

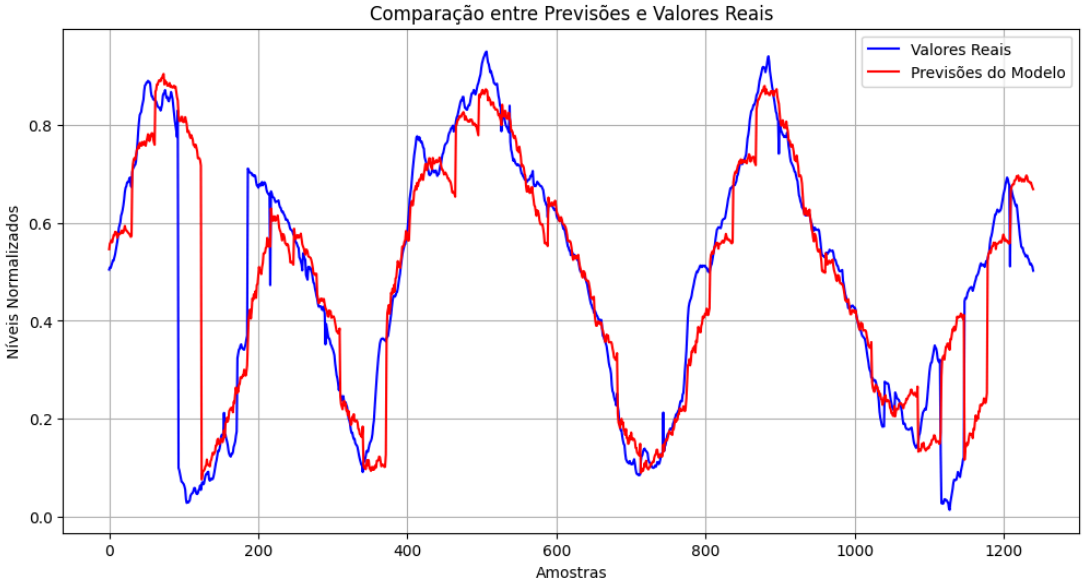


Fonte: Autor

- O valor de MAE é relativamente baixo (0.1318), indicando que, em média, o modelo comete erros de magnitude moderada nas suas previsões.
- O valor de MSE é ainda menor (0.0429), o que sugere que os erros do modelo estão concentrados em valores menores.
- O valor de RMSE (0.2072) é o maior entre as três métricas. Isso indica que, embora o modelo tenha um bom desempenho em termos de MAE e MSE, existem alguns casos em que os erros são consideravelmente maiores.

Estação Itaituba

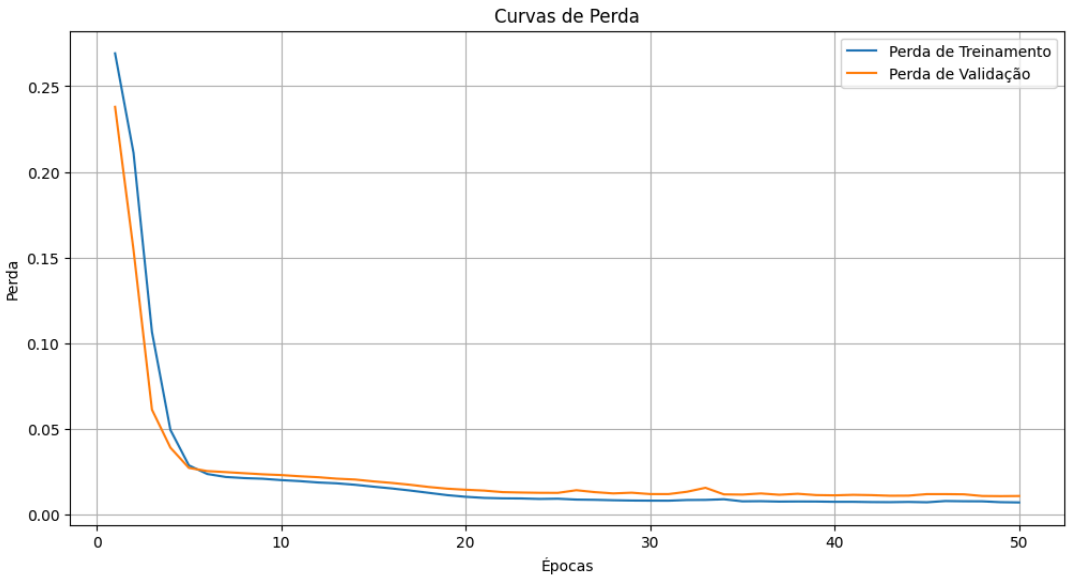
Figura 34 - Itaituba: Comparação de previsões com valores reais



Fonte: Autor

A análise gráfica indica que as predições do modelo conseguem acompanhar a tendência geral dos dados, o modelo converge bem em relação aos valores reais entre as amostras 200 a 1000 com desvios pontuais nas amostras de 0 a 200 e 1500 a 1200.

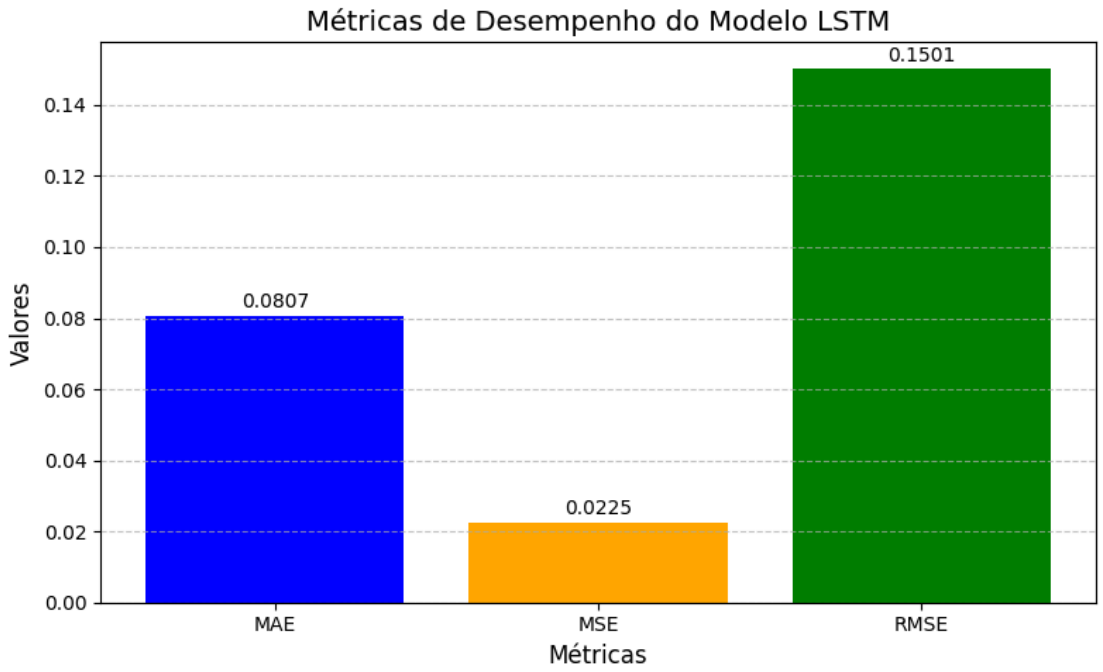
Figura 35 - Estação Itaituba: Função de perda no conjunto teste



Fonte: Autor

As curvas de perda diminuem significativamente nas primeiras épocas e então se estabilizam em um nível relativamente baixo. Isso sugere que o modelo está aprendendo os padrões dos dados de forma eficaz. A perda de validação acompanha de perto a perda de treinamento, indicando que o modelo está se generalizando bem para novos dados. Não há evidências claras de overfitting. O modelo parece ter convergido para um mínimo local, pois as curvas de perda se estabilizam após cerca de 30 épocas.

Figura 36 - Estação Itaituba: Resultados das métricas de desempenho

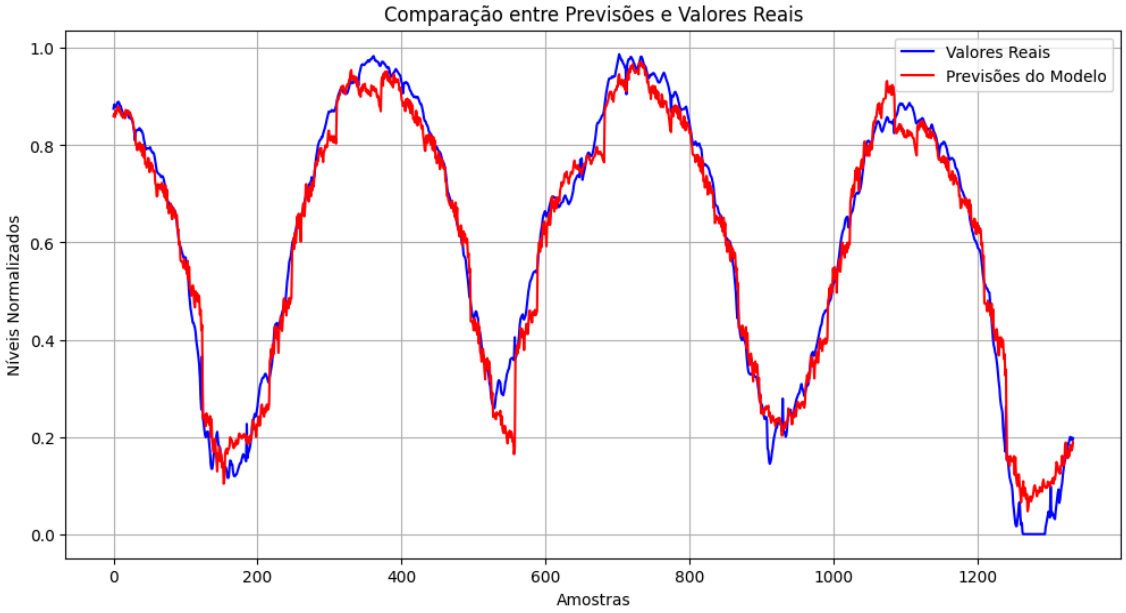


Fonte: Autor

- O valor de MAE é relativamente baixo (0.0807), indicando que, em média, o modelo comete erros de magnitude moderada nas suas previsões.
- O valor de MSE é ainda menor (0.0225), o que sugere que os erros do modelo estão concentrados em valores menores.
- O valor de RMSE (0.1501) é o maior entre as três métricas. Isso indica que, embora o modelo tenha um bom desempenho em termos de MAE e MSE, existem alguns casos em que os erros são consideravelmente maiores.

Estação Santarém

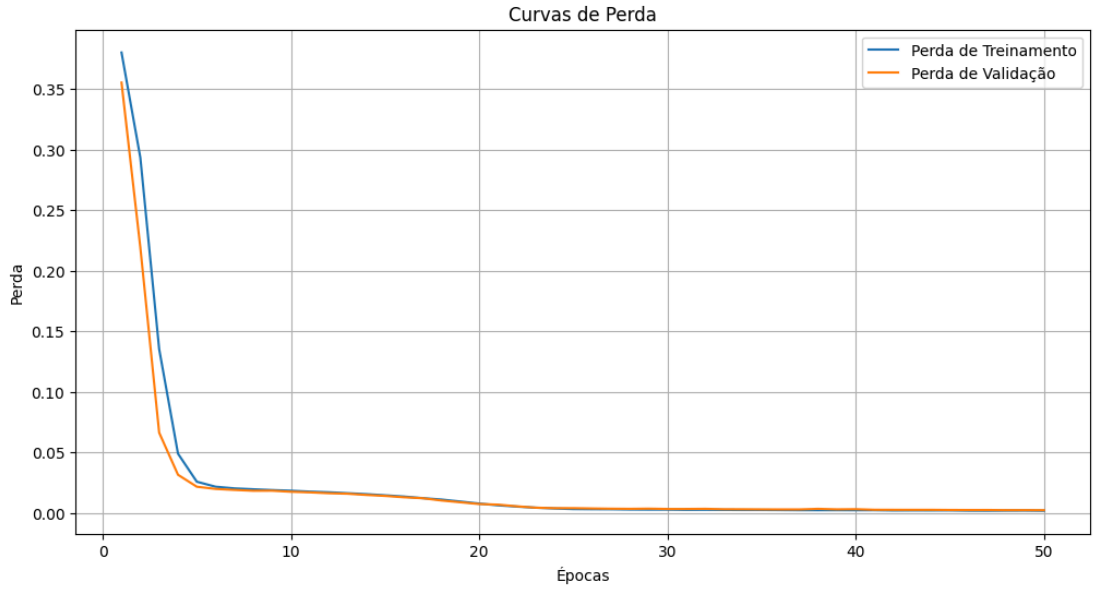
Figura 37 - Estação Santarém: Comparação de previsões com valores reais



Fonte: Autor

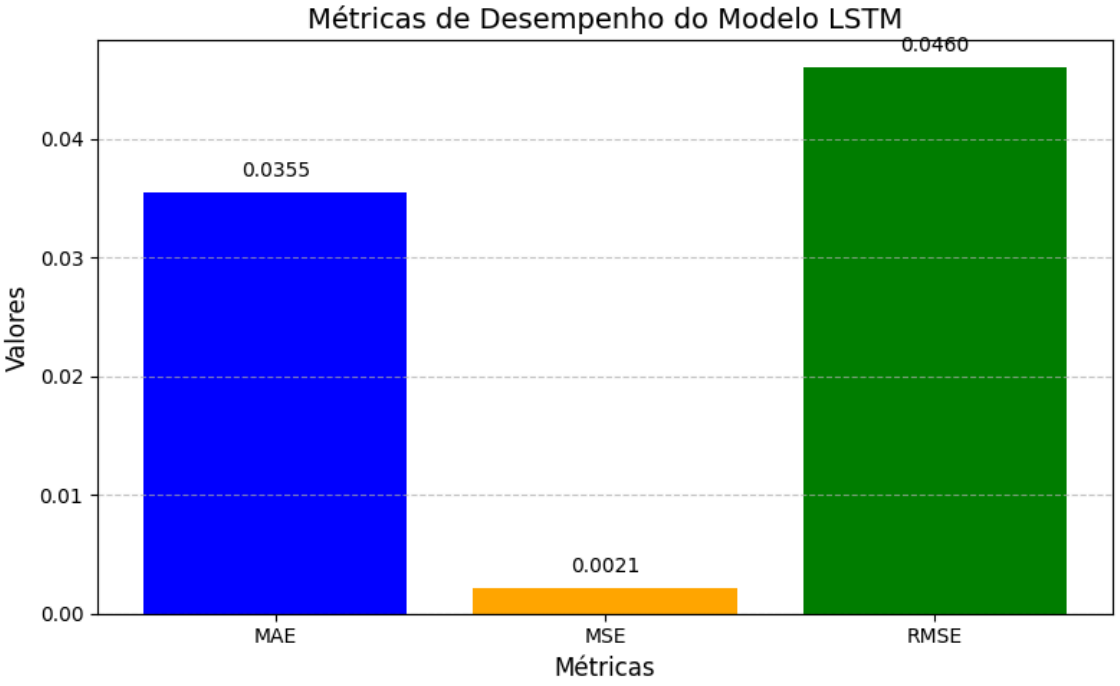
A análise gráfica indica que as predições do modelo conseguem acompanhar a tendência geral dos dados, o modelo converge bem em relação aos valores reais com exceção em alguns pontos específicos como os vales das amostras de 400 a 600, 800 a 100 e 1200 a 1300.

Figura 38 - Estação Santarém: Função de perda no conjunto teste



Fonte: Autor

Figura 39 - Estação Santarém: Resultados das métricas de desempenho



Fonte: Autor

- O valor de MAE é relativamente baixo (0.0355), indicando que, em média, o modelo comete erros de magnitude moderada nas suas previsões.
- O valor de MSE é ainda menor (0.0021), o que sugere que os erros do modelo estão concentrados em valores menores.
- O valor de RMSE (0.0460) é o maior entre as três métricas. Isso indica que, embora o modelo tenha um bom desempenho em termos de MAE e MSE, existem alguns casos em que os erros são consideravelmente maiores.

5.1. Predição autoregressiva

Nesta seção serão comparados os valores reais dos cincos primeiros meses de 2024 que não foram mostrados ao modelo com os valores preditos pelo modelo LSTM. Esse método realiza previsões futuras em uma sequência de dados temporais utilizando um modelo LSTM gerando os cinco meses preditos.

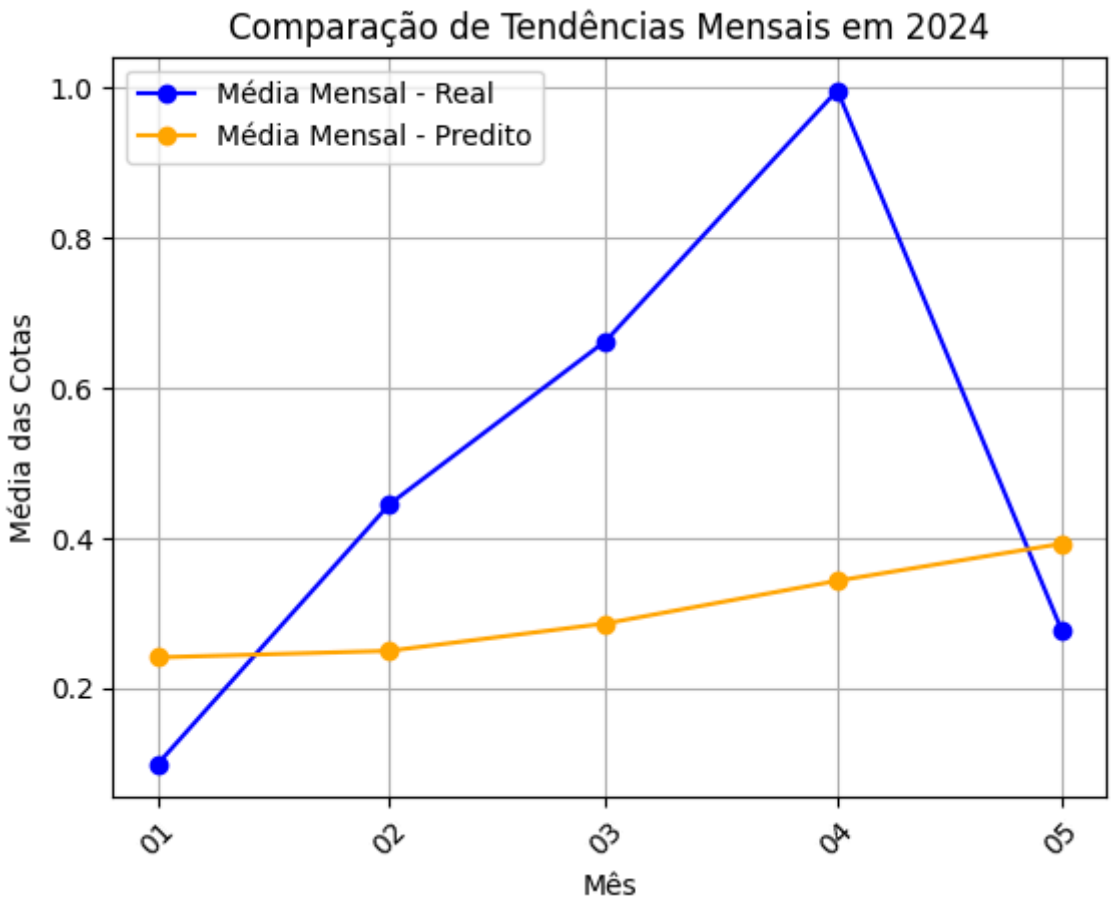
Utilizando a última sequência do conjunto de teste como base para a primeira previsão e realiza a iteração sobre de acordo com o valor escolhido (no caso 5), e

armazena os resultados em uma lista que posteriormente será transformada em um conjunto de dados para realizar as análises das predições.

Estação Barra do São Manuel

- Comparativo dos dados reais e preditos

Figura 40 - Estação Barra do São Manuel: Resultados reais e preditos de 2024



Fonte: Autor

Tabela 5 - Estação Barra do São Manuel: Resultados reais e preditos de 2024

Data	Média mensal Real	Média mensal Predições
2024-01-01	0.099500	0.241255
2024-02-01	0.445249	0.250032
2024-03-01	0.663059	0.286548

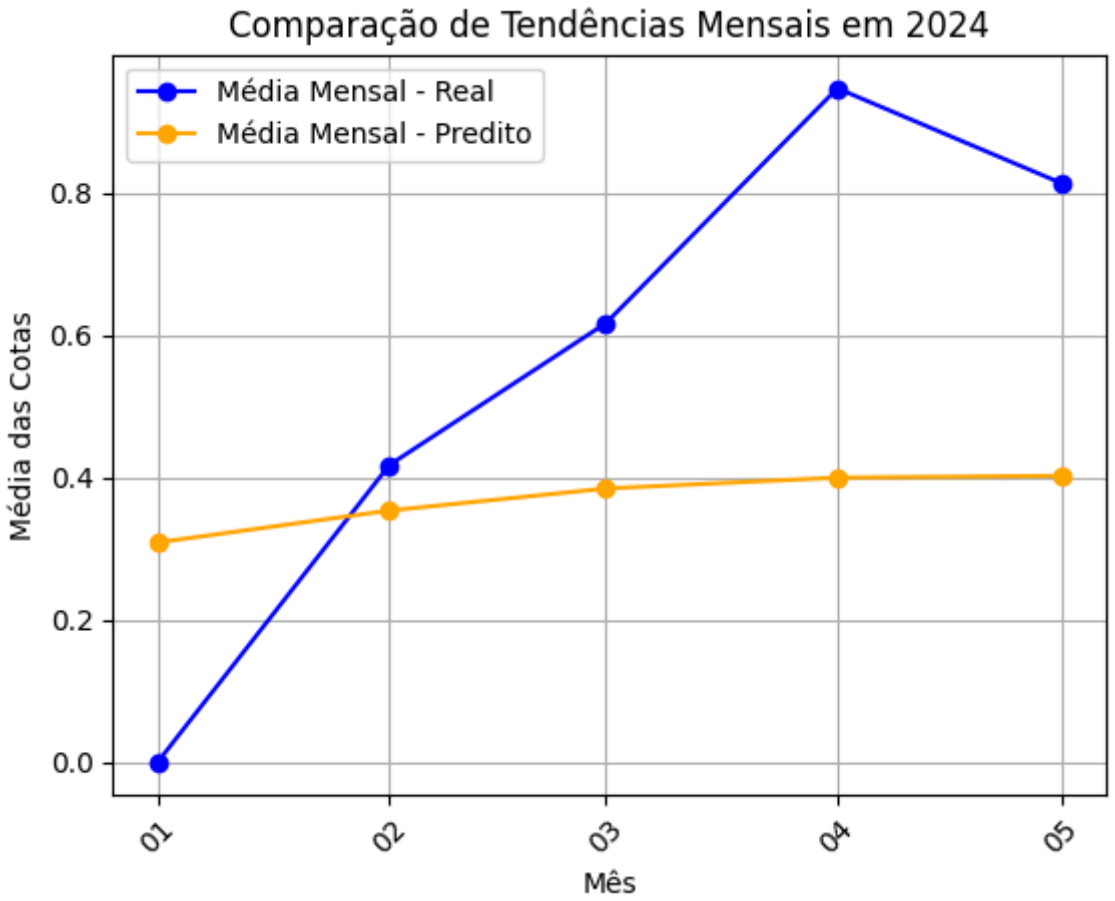
2024-04-01	0.996774	0.343364
2024-05-01	0.278521	0.392653

Fonte: Autor

Estação Itaituba

- Comparativo dos dados reais e preditos

Figura 41 - Estação Itaituba: Resultados reais e preditos de 2024



Fonte: Autor

Tabela 6 - Estação Itaituba: Resultados reais e preditos de 2024

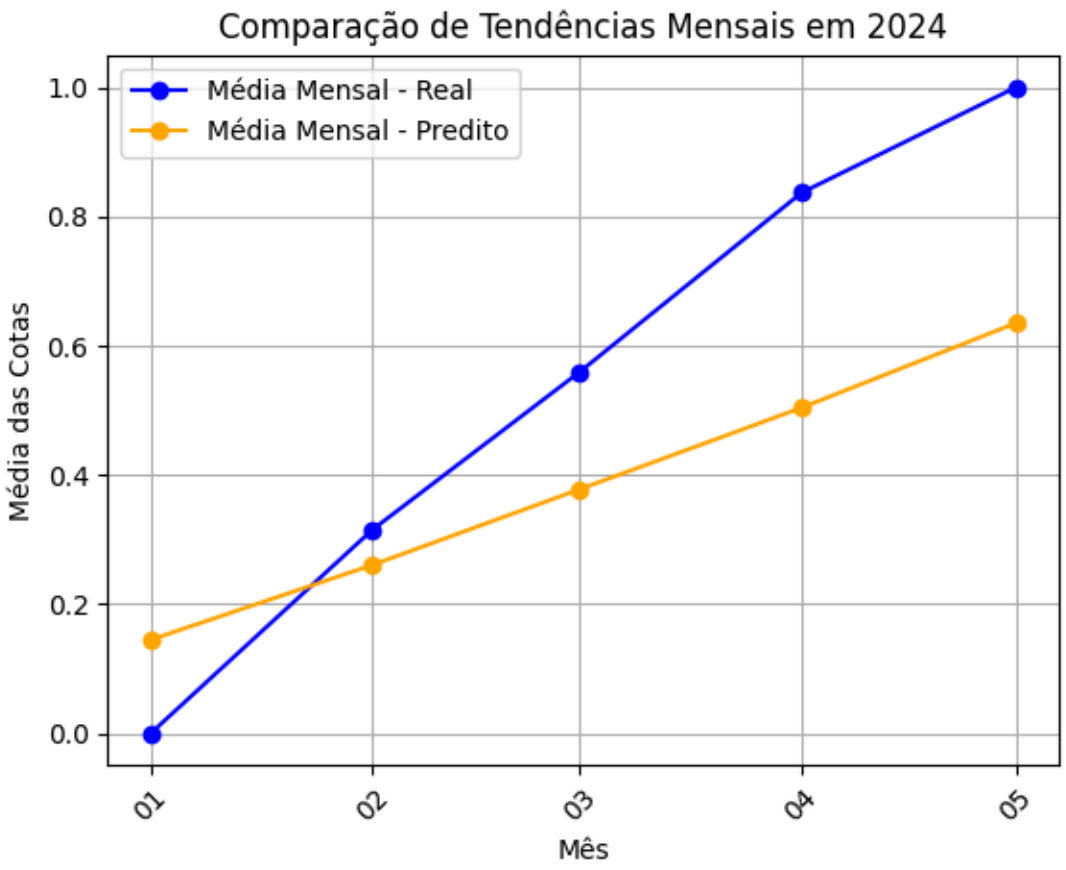
Data	Média mensal Real	Média mensal Predições
2024-01-01	0.099500	0.241255
2024-02-01	0.445249	0.250032
2024-03-01	0.663059	0.286548
2024-04-01	0.996774	0.343364
2024-05-01	0.278521	0.392653

Fonte: Autor

Estação Santarém

- Comparativo dos dados reais e preditos

Figura 42 - Estação Santarém: Resultados reais e preditos de 2024



Fonte: Autor

Tabela 7 - Estação Santarém: Médias mensais reais de 2024

Data	Média mensal Real	Média mensal Predições
2024-01-01	0.000000	0.144661
2024-02-01	0.315161	0.260509
2024-03-01	0.559580	0.377794
2024-04-01	0.836744	0.503900
2024-05-01	1.000000	0.635031

Fonte: Autor

5.2. Discussão dos resultados e conclusões

5.2.1. Treinamento e validação do modelo

Ao analisar o desempenho do modelo em relação aos dados das três estações hidroclimatológicas, é possível observar diferenças notáveis na capacidade de convergência do modelo entre elas. A estação de Santarém destacou-se com a melhor convergência, apresentando os resultados mais consistentes em termos de precisão das previsões. A estação Barra do São Manuel apresentou os piores resultados de convergência, com base nas reflexões dos valores das métricas de avaliação.

Fatores a Considerar:

Convergência do Modelo: Na estação de Santarém, o modelo atingiu uma convergência mais rápida e eficiente, refletida nos valores das métricas de avaliação (como erro quadrático médio). Isso sugere que os dados dessa estação podem ser mais regulares ou menos ruidosos, facilitando a identificação de padrões pelo modelo.

Diferenças entre as estações: As estações de Barra do São Manuel e Itaituba podem ter apresentado mais variabilidade ou ruído em seus dados, o que pode ter dificultado a convergência do modelo.

5.2.2. Predição autoregressiva

Os resultados das previsões autoregressivas do modelo LSTM seguem o padrão sazonal dos dados de níveis do rio, o que indica que o modelo conseguiu capturar e aprender o comportamento cíclico presente nos dados históricos. Essa observação é crucial, pois a capacidade de identificar padrões sazonais é uma característica essencial em problemas de previsão de séries temporais, especialmente em contextos hidrológicos, onde os níveis dos rios seguem flutuações sazonais bem definidas.

No entanto, o que difere nas previsões do modelo em comparação aos dados reais é a escala de proximidade. Embora o modelo tenha aprendido o comportamento geral e a forma do ciclo sazonal, as previsões geradas tendem a ser mais suavizadas ou apresentar uma amplitude menor em comparação aos valores observados nos dados reais. Isso sugere que, enquanto o modelo conseguiu prever corretamente os altos e baixos sazonais, a intensidade ou magnitude das variações pode não ter sido totalmente capturada.

6. CONTRIBUIÇÕES E PROPOSTA DE CONTINUAÇÃO DA PESQUISA

6.1. Contribuições da pesquisa

Este estudo aplicou o uso das redes neurais recorrentes para monitoramento hidrológico, especificamente para predição de níveis de cota com base em dados hidroclimatológicos disponíveis publicamente. A metodologia aplicada pode ser expandida para outras regiões e sistemas fluviais utilizando a mesma fonte de dados, demonstrando a versatilidade das RNNs no campo de predições ambientais.

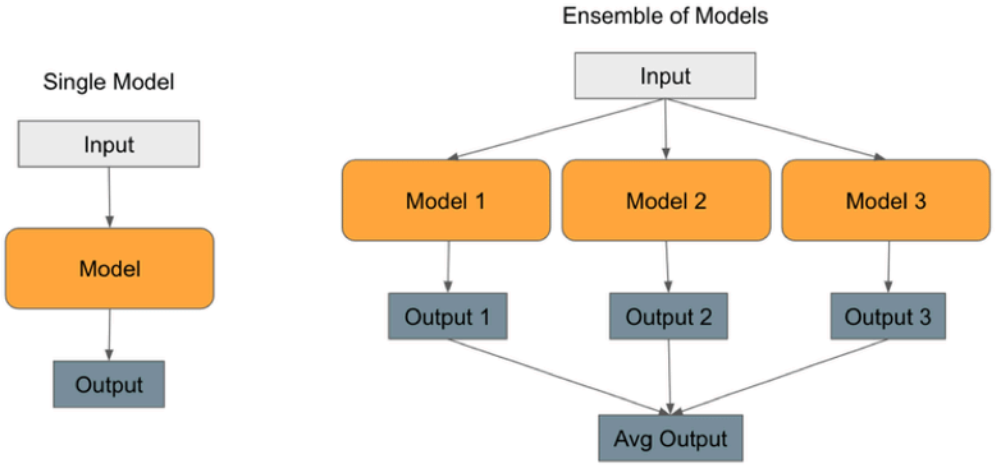
No contexto local, este estudo contribui ao explorar a aplicação de algoritmos de inteligência artificial no monitoramento ambiental do rio Tapajós. Ao demonstrar o uso de redes neurais para predição de níveis de cota, esta pesquisa abre a possibilidade para o desenvolvimento de soluções tecnológicas voltadas para a gestão de recursos hídricos. Além de fornecer uma ferramenta para auxiliar no monitoramento do rio.

6.2. Continuação da pesquisa

Model ensembling é uma técnica para obter os melhores resultados possíveis em uma tarefa. O ensembling consiste em reunir as previsões de um conjunto de modelos diferentes para produzir melhores previsões (CHOLLET, 2018).

Combinar o modelo LSTM com outras arquiteturas de como GRU (Gated Recurrent Units), redes neurais convolucionais (CNN) para extração de padrões em séries temporais ou utilizar modelos híbridos que combinam LSTM e CNNs podem levar a uma melhoria no desempenho e na capacidade de predição do modelo.

Figura 43 - Exemplo de model ensembling



Fonte:
https://www.researchgate.net/figure/An-illustration-of-ensembling-models-On-the-left-is-a-single-model-which-takes-input_fig16_332169081

Incluir variáveis adicionais, como dados de precipitação obtidos das estações fluviométricas e os dados de vazão, pode melhorar significativamente a precisão do modelo. Obtendo essas observações e combinando modelos é possível fazer uma correlação de chuva-cota-vazão para convergir em um resultado mais preciso. A análise de variáveis climáticas regionais poderia fornecer uma compreensão mais profunda das relações entre o clima e as variações nos níveis do rio.

7. REFERÊNCIAS

Dive into Deep Learning — Dive into Deep Learning 1.0.0-alpha1.post0 documentation. Disponível em: <https://d2l.ai>. Acesso em: 5 nov. 2024.

CHOLLET, F. Deep Learning with Python. Shelter Island (New York, Estados Unidos): Manning, Cop, 2018.

DANIEL GOMES FERRARI; NUNES, L. Introdução à mineração de dados. [s.l.] Saraiva Educação S.A., 2017.

SILVA. Introdução à Mineração de Dados - Com Aplicações em R. [s.l: s.n.].

WOLF, Andrew. *The Machine Learning Simplified: A Gentle Introduction to Supervised Learning*. Edição em inglês. [eBook Kindle]. 2022.

MINÉRIO, Daniel. Introdução à Mineração de Dados - Com Aplicações em R. 1. ed. São Paulo: Editora Ciência Moderna, 2016.

AGÊNCIA NACIONAL DE ÁGUAS – ANA. Manual de Procedimentos Operacionais para Estações de Monitoramento Hidrometeorológico. Brasília: ANA, 2014. Disponível em: https://www.gov.br/ana/pt-br. Acesso em: 5 out. 2024.

AGÊNCIA NACIONAL DE ÁGUAS. Manual de Procedimentos para Instalação, Operação e Manutenção de Estações Fluviométricas. Brasília: ANA, 2016. Acesso em: 5 out. 2024.

FERNANDES, Vicente; GUERRA, Antônio. Dinâmica fluvial e gestão dos recursos hídricos. Rio de Janeiro: EdUERJ, 2016.

LINSLEY, Ray K.; KOHLER, Max A.; PAULHUS, Joseph L. H. Hydrology for Engineers. 3. ed. New York: McGraw-Hill, 1982.

TUCCI, Carlos E. M. Hidrologia: Ciência e Aplicação. Porto Alegre: Ed. Universidade/UFRGS, 2009.

BENGIO, Yoshua et al. Learning Long-Term Dependencies with Gradient Descent is Difficult. IEEE Transactions on Neural Networks, v. 5, n. 2, p. 157-166, 1994.

CHO, Kyunghyun et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078, 2014.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep Learning. Cambridge: MIT Press, 2016.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. Neural Computation, v. 9, n. 8, p. 1735-1780, 1997.

KRATZERT, Frederik et al. Towards Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resources Research, v. 55, n. 12, p. 11344-11354, 2019.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 1996.

HAN, J., KAMBER, M. Data Mining: Concepts and Techniques. 3rd Edition, Morgan Kaufmann, 2012.

WITTEN, I. H., FRANK, E., HALL, M. A. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition, Elsevier, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Site da Hidroweb disponível em: https://www.snirh.gov.br/hidroweb/apresentacao

Estações de hidro-telemetria disponível em: https://www.snirh.gov.br/hidrotelemetria

streamflow_tapajos_data_pipeline.ipynb disponível em: https://colab.research.google.com/drive/12siUSRanoAnv0LWYxZafXtN1oC8BbDZv#scrollTo=TIwmRN9xIozS

lstm_pipeline.ipynb disponível em: https://colab.research.google.com/drive/1pPoW6NchEhdl5pPzceKhnOaljk9yo8Rb#scrollTo=F8Qb72HYNHA0