UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Gabryel Martins Raposo de Alencar

Aplicação de Soft Sensor para Macromedição de vazão utilizando Redes Neurais Artificiais

João Pessoa - PB
2023

Gabryel Martins Raposo de Alencar

Aplicação de Soft Sensor para Macromedição de vazão utilizando Redes Neurais Artificiais

Trabalho de Conclusão de Curso apresentado à Coordenação de Engenharia Elétrica do Centro de Energias Alternativas e Renováveis da Universidade Federal da Paraíba como parte dos requisitos necessários para a obtenção do título de Engenheiro Eletricista.

Universidade Federal da Paraíba

Centro de Energias Alternativas e Renováveis

Curso de Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Juan Moises Mauricio Villanueva

João Pessoa - PB

2023

Catalogação na publicação
Seção de Catalogação e Classificação

A368a Alencar, Gabryel Martins Raposo de.
Aplicação de Soft Sensor para Macromedição de vazão utilizando Redes Neurais Artificiais / Gabryel Martins Raposo de Alencar. - João Pessoa, 2023.
44 f. : il.

Orientação: Prof Dr Juan Moises Mauricio Villanueva.
TCC (Graduação) - UFPB/CEAR.

1. Soft Sensor. 2. Redes Neurais Artificiais. 3. LSTM. 4. Indústria 4.0. 5. TEDA. 6. Perdas de Água. 7. Sistemas de Abastecimento. I. Villanueva, Prof Dr Juan Moises Mauricio. II. Título.

UFPB/CT/BSCT CDU 621.3(043.2)

Elaborado por ONEIDA DIAS DE PONTES - CRB-15/198

Gabryel Martins Raposo de Alencar

Aplicação de Soft Sensor para Macromedição de vazão utilizando Redes Neurais Artificiais

Trabalho de Conclusão de Curso apresentado à Coordenação de Engenharia Elétrica do Centro de Energias Alternativas e Renováveis da Universidade Federal da Paraíba como parte dos requisitos necessários para a obtenção do título de Engenheiro Eletricista.

Data de Aprovação: 31 / 10 / 2023

Prof. Dr. Juan Moises Mauricio Villanueva
(Orientador)
Universidade Federal da Paraíba

Prof. Dr. Cícero da Rocha Souto
(Avaliador)
Universidade Federal da Paraíba

Dr. Rafael Moura Duarte
(Avaliador)
Universidade Federal da Paraíba

Documento assinado digitalmente
gov.br ALTAMAR ALENCAR CARDOSO
Data: 03/11/2023 12:00:23-0300
Verifique em https://validar.iti.gov.br

MSc. Altamar Alencar Cardoso
(Avaliador)
CAGEPA

João Pessoa - PB
2023

Dedico este trabalho a Jesus e Nossa Senhora.

Agradecimentos

Dedico este trabalho com profunda gratidão a Deus, por sua orientação e bênçãos que tornaram possível esta jornada.

Agradeço primeiramente a minha família. À minha mãe, Renata, pela força inabalável e por nunca deixar faltar nada, ao meu irmão, Raphael, pelo companheirismo e ao meu pai Arilo pela inspiração. Aos avós, Elsa e Severino, meu carinho e gratidão por moldarem meu caráter e aos padrinhos, Cristiano e Danyele, dedico este trabalho com carinho pelo apoio e incentivação.

Minha namorada, Letícia, agradeço por seu apoio incondicional e por todos os momentos juntos durante esses anos.

Aos amigos de infância, Alexandre Brindeiro e Pedro Henrique, pelo vínculo de longa data, meu agradecimento por todos os momentos vivenciados que me fortaleceram como pessoa e profissional.

Também deixo os agradecimentos aos amigos da graduação como Everlan Santana, Elton Davi, Maria Heloísa, Ruben da Cruz, Pedro Ravel que permaneceram ao meu lado na faculdade, nos momentos difíceis e que contribuíram profundamente na minha caminhada. Aos amigos Pedro Coutinho e Davi Ferreira, mesmo seguindo outros caminhos, foram uma fonte de força inicial.

Minha gratidão ao professor orientador, Prof. Dr. Juan Maurício, pelas orientações valiosas, ensinamentos repassados e solicitude e ao Prof. Dr. Alexandre Castro agradeço pela paciência e apoio durante todo o curso.

Agradeço a instituição e equipe da Cagepa pela confiança depositada em mim, assim como ao CNPQ pelo apoio na chamada CNPQ/MCTI/SEMPI Nº 56/2022 - Apoio para estudante elaborando TCC em Inteligência Artificial.

À instituição de ensino e aos professores que enriqueceram minha jornada, meu mais sincero agradecimento.

Finalmente, minha gratidão se estende a todos que contribuíram para a minha formação, desde os professores do ensino fundamental até os dias de hoje.

"O coração do homem planeja o seu caminho, mas o Senhor lhe dirige os passos."
(Provérbios 16,9)

Resumo

A quarta revolução industrial provocou uma transformação significativa na indústria, impulsionada pela sinergia com tecnologia da informação que desempenha um papel crucial nessa mudança. A digitalização crescente dos sistemas industriais exige formas eficientes de sensoriamento e controle, e é nesse cenário que surgem os *Soft Sensors*, que têm o potencial de substituir sensores físicos convencionais, reduzindo custos e melhorando a eficiência. Com a Indústria 4.0, técnicas de estimação baseadas em *software*, como Redes Neurais Artificiais, oferecem uma alternativa eficaz. Este estudo explora a implementação de um modelo de *Soft Sensor* em um sistemas de abastecimento de água para realizar a predição da vazão no sistema. A implementação é feita com Redes Neurais do tipo LSTM e a análise de incertezas com o método de Monte Carlo Dropout. Os resultados deste estudo demonstram que os *Soft Sensor* podem prever a vazão de forma precisa, contribuindo para a redução das perdas de água e economia de custos. Esta abordagem pode ser uma solução valiosa para as concessionárias de água enfrentarem o desafio das perdas e garantirem um uso eficiente desse recurso vital.

Palavras-chave:*Soft Sensor*, Redes Neurais Artificiais, LSTM, Perdas de Água, Sistemas de Abastecimento, Indústria 4.0, TEDA.

Abstract

The fourth industrial revolution has brought about a significant transformation in the industry, driven by the synergy with information technology, which plays a crucial role in this change. The increasing digitization of industrial systems demands efficient sensing and control methods, giving rise to soft sensors that have the potential to replace traditional physical sensors, reducing costs and enhancing efficiency. With Industry 4.0, software-based estimation techniques, such as Artificial Neural Networks, provide an effective alternative.This study explores the implementation of a soft sensor model in a water supply system to predict flow rates within the system. The implementation is done using Long Short-Term Memory (LSTM) neural networks, with uncertainty analysis performed through the Monte Carlo Dropout method.The results of this study demonstrate that soft sensors can predict flow rates with precision, contributing to the reduction of water losses and cost savings. This approach offers a valuable solution for water utilities to address the challenge of minimizing losses and ensuring efficient use of this vital resource.

Keywords: Soft Sensors, Artificial Neural Networks, LSTM, Water Loss, Supply Systems, Industry 4.0, TEDA.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CAGEPA	Companhia de Água e Esgotos da Paraíba
IA	Inteligência Artificial
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MSE	*Mean Squared Error*
RMSE	*Root Mean Squared Error*
RNA	Redes Neurais Artificiais
RNR	Redes Neurais Recorrentes
NARX	*Nonlinear AutoRegressive with eXogenous inputs*
SAA	Sistema de Abastecimento de Água
SCADA	*Supervisory Control And Data Acquisition*
SS	*Soft Sensors*
SV	Sensores Virtuais
TEDA	*Tipicity and Eccentricity Data Analysis*

Sumário

1 Introdução

A gestão dos Sistemas de Abastecimento de Água (SAA) torna-se cada vez mais um problema desafiador, impulsionado pela confluência da crescente demanda por recursos hídricos e a necessidade de eficiência energética, dada a considerável energia requerida para operações de bombeamento. A otimização desses sistemas assume um papel importante não apenas na garantia da qualidade e quantidade da água distribuída, mas também na mitigação do impacto ambiental e na redução dos custos operacionais.

Nesse panorama, os sistemas de abastecimento de água se configuram como um território vital para a convergência entre tecnologia e sustentabilidade, demandando uma abordagem inovadora que abranja tanto as necessidades humanas quanto os imperativos ambientais. A medição precisa da vazão emerge como um elemento crítico nessa equação, possibilitando o monitoramento contínuo das operações, a identificação de perdas e vazamentos, e a adaptação dinâmica às flutuações de demanda.

A complexidade intrínseca aos sistemas de abastecimento de água, envolvendo componentes interdependentes, amplia o desafio da medição da vazão. Aqui, a inteligência artificial (IA) e os princípios da Indústria 4.0 se destacam como ferramentas promissoras para aprimorar o controle e medição desses sistemas. Os sensores virtuais, ou *soft sensors* (SS), exemplificam essa conjunção, combinando a coleta e análise de dados em tempo real com algoritmos avançados de IA para estimar a vazão com base em informações indiretas.

A utilização da IA oferece diversas vantagens, sendo aplicável em empresas de saneamento que enfrentam desafios relacionados à qualidade de seus dados e possuem recursos limitados tanto para aquisição de equipamentos de medição quanto para o aprimoramento da capacitação de suas equipes técnicas (ROCHA, 2018).

Contudo, a aplicação da IA e da Indústria 4.0 para otimizar os SAA não é isenta de desafios. A concepção de algoritmos de IA exige uma compreensão aprofundada das operações dos SAA, além da tradução dessas percepções em indicadores operacionais tangíveis para gerentes e operadores. Além disso, a diversidade entre sistemas e as limitações de medição podem dificultar a criação de indicadores de eficiência energética aplicáveis universalmente.

Segundo o Sistema Nacional de Informações sobre Saneamento (SNIS), as perdas de distribuição de água no Brasil foram de 40,25% e, em âmbito do Estado da Paraíba, essas perdas caem para 35,38% (SNIS, 2021). Sendo assim, tais perdas representam um problema significativo que pode ser mitigado por meio da medição precisa da vazão, contribuindo para a eficiência operacional dos Sistemas de Abastecimento de Água.

É fundamental destacar que o Brasil possui um Marco Legal do Saneamento Básico regido pelo Projeto de Lei n° 4.162/2019, em que uma das metas proeminentes é que até dezembro de 2033, 99% da população brasileira deverá ter acesso à água potável em suas residências (LIMA, 2022b). Em um contexto mais atual, o marco regulatório de saneamento em vigência, aprovado em 2020, estabelece regras para as concessionárias de saneamento, incluindo a redução progressiva e o controle das perdas de água, bem como metas de universalização dos serviços públicos (BRASIL, 2020).

Além disso, parte significativa do marco legal está relacionada ao incentivo ao investimento no setor de saneamento. Esses investimentos incluem não apenas financiamento para infraestrutura, mas também o aprimoramento de novas metodologias, como a utilização de IA, para combater as perdas de água.

Assim, este Trabalho de Conclusão de Curso procura convergir essas diversas dimensões discutidas ao empregar a abordagem dos *soft sensors*, fundamentados em algoritmos de IA e nos princípios da Indústria 4.0, almejando superar as limitações inerentes às medições convencionais e criando uma base sólida para a otimização operacional dos SAA realizando a predição de vazão de água. Para tanto, será realizado um estudo de caso com dados da Companhia de Água e Esgotos da Paraíba (CAGEPA). Por meio dessa abordagem integrativa, visamos avançar não apenas na medição da vazão em sistemas de abastecimento de água, mas também contribuir para uma gestão mais sustentável, eficiente e informada dos recursos hídricos.

1.1 Objetivos

O escopo central deste trabalho é o desenvolvimento de uma metodologia para a construção de um instrumento virtual avançado, denominado sensor virtual ou *soft sensor*, com a finalidade primordial de estimar a vazão em um sistema de abastecimento de água, utilizando de Redes Neurais Artificiais. Entre os objetivos específicos estão:

- Redução substancial dos custos associados à obtenção de medidores físicos de vazão;
- Simplificação da complexidade relacionada à instalação desses dispositivos em locais específicos da infraestrutura.

1.2 Organização do trabalho

Dentro deste estudo, para além deste primeiro capítulo introdutório, estão dispostos mais quatro capítulos que serão apresentados a seguir.

No segundo capítulo, está reservado o espaço para a exposição da base teórica a respeito do tópico abordado. Neste capítulo, aprofundam-se os conceitos envolvendo os

Soft Sensors, ao mesmo tempo em que se realiza uma análise mais detalhada sobre Redes Neurais Artificiais(RNAs) e a consideração de incertezas no âmbito da medição.

O terceiro capítulo engloba a exposição do estudo de caso, onde são fornecidas as informações cruciais sobre a planta industrial e seus respectivos processos. Também é descrito o banco de dados e delineado o sistema empregado para a coleta dos dados. Além disso, neste mesmo capítulo, é dedicado espaço para a introdução aos métodos de seleção de variáveis, um passo crucial para o treinamento dos modelos preditivos. E, concluindo o terceiro capítulo, são apresentados os algoritmos para a estimativa da vazão, bem como a abordagem de estimação de incertezas, além das métricas que foram utilizadas para avaliar o desempenho dos modelos.

O quarto capítulo é destinado à exposição da implementação prática e dos resultados obtidos. Aqui, são apresentados os modelos desenvolvidos para a estimação de vazão, considerando as particularidades do tratamento de dados e a remoção de dados discrepantes (*outliers*), assim como o treinamento de diversas redes neurais. Ainda são detalhadas as métricas de avaliação utilizadas e realizadas comparações entre os diferentes modelos desenvolvidos. Adicionalmente, é apresentada a estimação de incerteza associada ao melhor modelo obtido.

Por fim, no quinto capítulo, é reservado o espaço para as considerações finais. Este é o espaço onde se apresentam as conclusões do estudo, juntamente com as perspectivas para trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo, são apresentados os conceitos-chave que servirão de base para o desenvolvimento deste trabalho. Inicialmente, é abordado o funcionamento dos Sistemas de Abastecimento de Água (Seção 2.1) e, em seguida, discutido os princípios da Inteligência Artificial (Seção 2.2). Na sequência, as Redes Neurais Artificiais (Seção 2.3) e, mais especificamente, as Redes Neurais LSTM (*Long Short-Term Memory*) (Seção 2.4). Também é introduzido o conceito do TEDA (*Tipicity and Eccentricity Data Analytics*) (Seção 2.6) e sua aplicação na detecção de *outliers* (Seção 2.6.1). Por fim, são discutidas as métricas de avaliação dos modelos (Seção 2.8), incerteza e Monte Carlo Dropout (Seção 2.8).

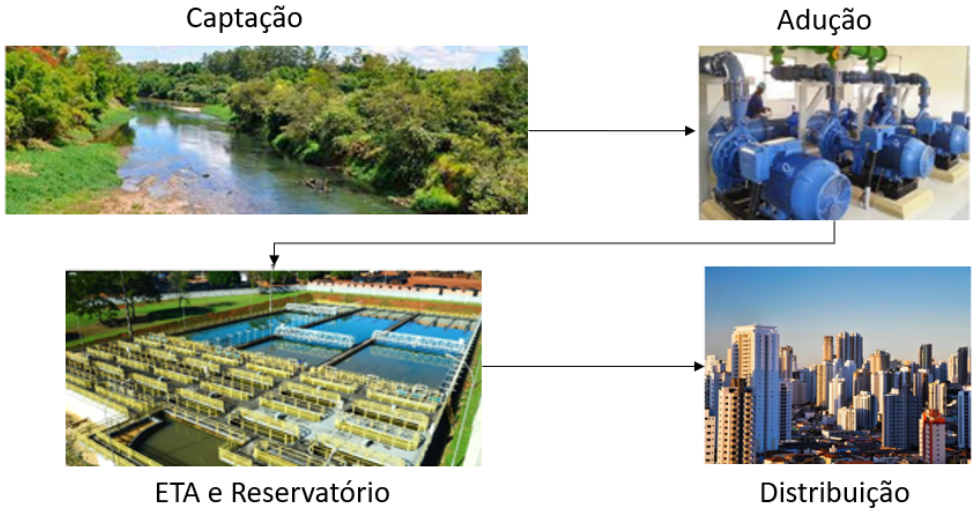
2.1 Sistema de abastecimento de Água

Um SAA se refere ao processo completo de abastecimento de água desde a sua captação até recebimento nas casas da sociedade. Desse modo, as principais etapas de um SAA são descritas a seguir:

- **Captação de Água**: A primeira etapa envolve a identificação e captação da água bruta de fontes naturais, como rios e lagos, por meio de equipamentos (bombas) e direcionar a água para a próxima fase do processo;
- **Adução da Água Bruta e Estações Elevatórias**: Após a captação, a água bruta é conduzida por meio de estações elevatórias e tubulações até a próxima etapa. Nesse processo ocorre o transporte da água das fontes de captação para a Estação de Tratamento de Água;
- **Estação de Tratamento de Água (ETA)**:A água bruta captada passa por um processo de tratamento em uma ETA. O tratamento envolve várias etapas, como coagulação, floculação, decantação, filtração e desinfecção. Durante esses processos, impurezas, sólidos suspensos, microorganismos patogênicos e substâncias químicas indesejadas são removidos ou neutralizados. Ajustes no pH e adição de produtos químicos são realizados para garantir que a água atenda aos padrões de qualidade e seja segura para o consumo humano;
- **Reservatório**: A água tratada é armazenada em reservatórios ou torres de água que servem como tambores de armazenamento temporário, permitindo que a água seja distribuída de forma contínua, mesmo quando a demanda varia ao longo do dia;
- **Distribuição**: A água do reservatório, após passar por tratamento, é distribuida para a sociedade em suas residências nas pequenas cidades e metrópoles.

A ilustração na Figura 1 representa o processo desse sistema:

Figura 1 – Ilustração de um exemplo de Sistema de Abastecimento de Água.



Fonte: Autoria Própria.

Durante todo o processo de captação e distribuição da água, sensores de pressão, nível e vazão são instalados para realizar o controle e monitoramento do processo. Atualmente, na Companhia de Água e Esgotos da Paraíba (CAGEPA), a utilização de sensores físicos convencionais para medir a vazão é prática comum. No entanto, o foco central deste trabalho de conclusão de curso é propor uma abordagem inovadora, que consiste em estimar a vazão a partir das variáveis conhecidas, como pressão e nível da água. Essa abordagem oferece torna desnecessário a aquisição de sensores de vazão convencionais, o que não apenas resulta em economia de recursos considerável, mas também elimina a necessidade de instalação física desses sensores, simplificando significativamente o processo de medição.

2.2 Inteligência Artificial

De acordo com (RUSSELL; NORVIG, 2010), a Inteligência Artificial (IA) é o campo interdisciplinar em constante expansão que faz uso de algoritmos e modelos computacionais avançados para simular processos cognitivos humanos. Quando atinge seu auge, essa capacidade permite que a IA tome decisões de maneira racional, aproximando-se das habilidades cognitivas humanas. Esse contínuo desenvolvimento da IA resulta em avanços notáveis na automação e na solução de problemas complexos, como predição de variáveis e parâmetros críticos em processos industriais contribuindo para otimização de sistemas de controle.

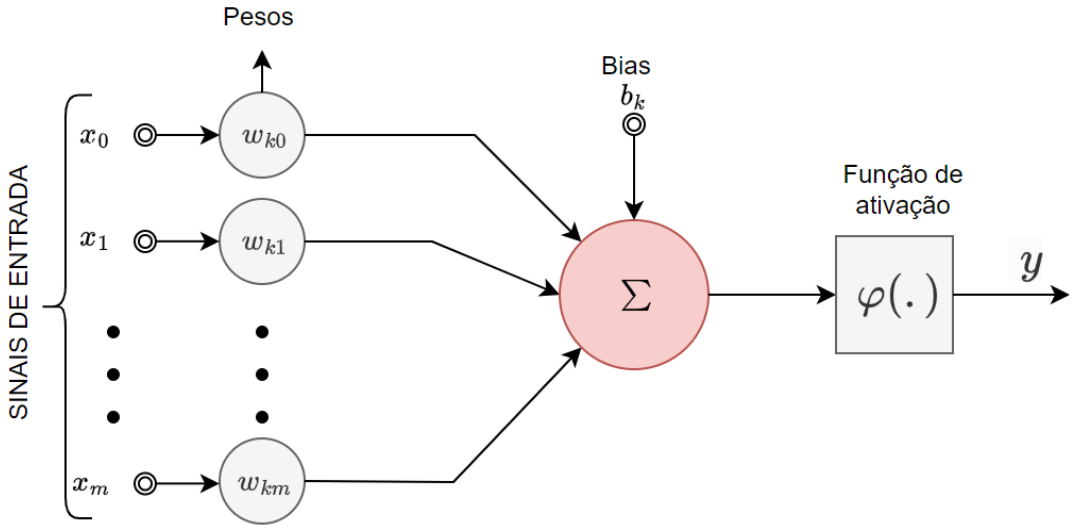
É importante mencionar conceitos essenciais, como o Aprendizado de Máquina (*Machine Learning*), que inclui métodos supervisionados, não supervisionados e por reforço. O *Machine Learning* (ML) capacita sistemas a aprenderem com dados, aprimorando seu desempenho com o tempo e buscando padrões e correlações entre eles *(SCHLEDER et al., 2019)*. Uma subcategoria do ML, o *Deep Learning*, se destaca pelo uso de RNAs para extrair representações complexas de dados, encontrando aplicações significativas em visão computacional, na criação de sensores virtuais, entre outros.

Ainda no âmbito da IA, um novo conceito vem tornando forma, o *Explainable Artificial Intelligence* (XIA), é uma faceta essencial da Inteligência Artificial que diz respeito à capacidade de revelar e comunicar como as decisões são alcançadas dentro dos sistemas de IA. À medida que a IA se torna cada vez mais prevalente em aplicações críticas, como medicina, finanças e até mesmo controle de processos industriais, a transparência nas decisões tomadas por algoritmos de IA torna-se vital. A XIA visa trazer à tona o motivo por trás das decisões, desmembrando os tipos de modelo caixa preta (*black box*) tradicionais e entendendo o processamento dentro delas. Isso desempenha um papel crucial na construção da confiança nas tecnologias de IA, especialmente quando as decisões têm implicações significativas na vida das pessoas (ALGOLIA, 2023).

2.3 Redes Neurais Artificiais

Redes Neurais Artificiais, usualmente denominadas redes neurais, são modelos matemáticos que se assemelham às estruturas neurais biológicas (SOARES; SILVA, 2011), notadamente do cérebro humano. Similar aos neurônios biológicos, as RNAs são compostas por unidades fundamentais chamadas neurônios artificiais. Cada neurônio artificial é composto por dendritos para receber sinais de entrada e um axônio para transmitir informações processadas. O funcionamento das RNAs envolve a absorção de sinais de entrada, o processamento e a produção de uma saída, através de uma função de ativação. As conexões entre os neurônios artificiais possuem pesos numéricos que determinam a importância das entradas para o cálculo da saída e seu processo de treinamento envolve ajustar esses pesos para que a rede possa aprender a realizar tarefas específicas. Desse modo, as RNAs são projetadas para realizar tarefas complexas de aprendizado de máquina e processamento de informações, com aplicação em diversas áreas. Por meio do diagrama da Figura 2, ilustra-se a configuração de um neurônio artificial.

Figura 2 – Modelo não linear de um neurônio artificial.



Fonte: Autoria Própria.

2.4 Redes Neurais Recorrentes

As Redes Neurais Recorrentes (RNR) são redes neurais que utilizam dados sequenciais, comumente implementados em problemas de séries temporais por possuirem conexões de retroalimentação, permitindo que sejam propagados dados de processamentos passados aos atuais (PEREIRA, 2023). Uma forma mais simples de entendimento das Redes Recorrentes seria pensar em uma espécie de memória da célula, que captura informações de etapas anteriores e perpassa as próximas etapas. Entretanto, tais redes tradicionais, também conhecidas como RNR *vanilla*, sofrem um problema de esquecimento do gradiente ao tentar aprender conexões de longo prazo.

O gradiente é definido como o valor que atualiza os pesos da rede neural e o problema do esquecimento acontece quando o gradiente se torna pequeno ou desaparece durante o processo do *backpropagation* de forma a não contribuir com o aprendizado da rede (MATSUMOTO; DUARTE; MURAKAMI, 2019).

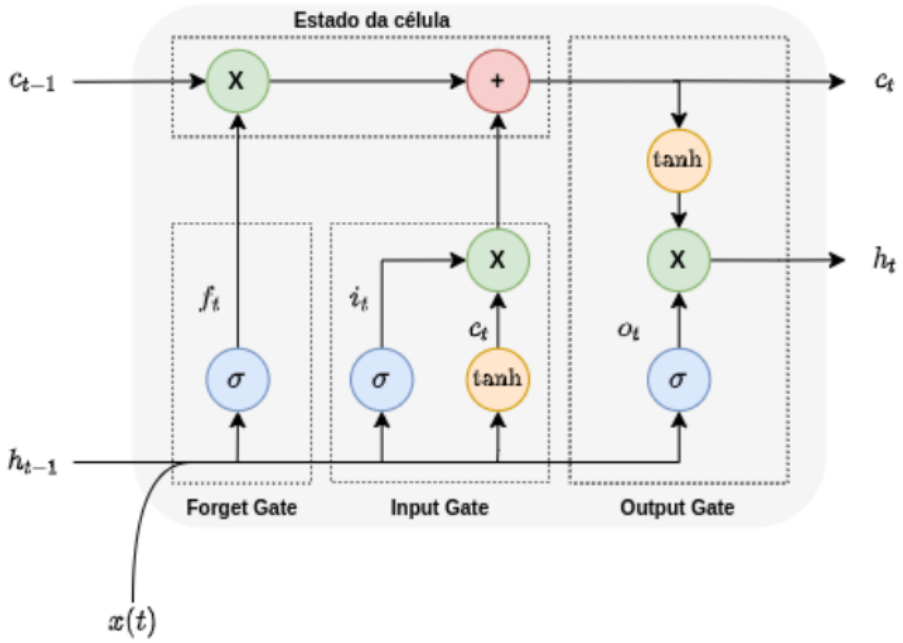
Dessa forma, surgem as Redes Neurais Recorrentes LSTM que têm sido amplamente adotadas em problemas de séries temporais devido à sua notável capacidade de lidar com o problema do gradiente. Diferentemente das redes neurais tradicionais, as LSTMs mitigam o desaparecimento do gradiente, permitindo que informações relevantes sejam mantidas por longos períodos, o que é crucial para a modelagem de séries temporais complexas e na modelagem de *soft sensors*.

Assim, a base de desenvolvimento do SS neste trabalho de conclusão de curso são as Redes Neurais do tipo LSTM.

2.5 Redes Neurais LSTM

As Redes Neurais *Long Short-Term Memory*, introduzidas por Hochreiter e Schmidhuber em 1997, representam um avanço significativo na área de Aprendizado Profundo, particularmente no processamento de sequências temporais. Uma característica distintiva das LSTMs é a presença de três portas fundamentais em cada neurônio: a porta de entrada (*input gate*), a porta de saída (*output gate*) e a porta de esquecimento (*forget gate*). Como está ilustrado na Figura 3:

Figura 3 – Neurônio Artificial LSTM.



Fonte: *(BARBOSA et al., 2021).*

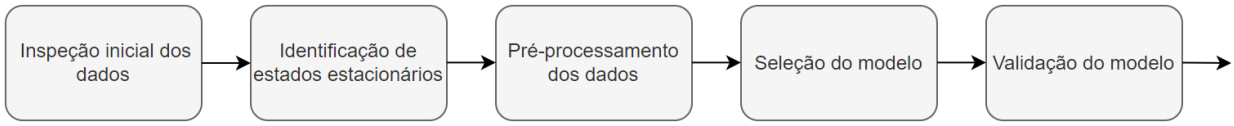
Essas portas permitem que as LSTMs controlem o fluxo de informações dentro da rede, o que as torna especialmente adequadas para capturar dependências temporais de longo prazo.

2.6 *Soft Sensors*

Conceitualmente um *soft sensor* é um modelo matemático, implementado em *software* e utilizado para estimar variáveis de interesse de difícil medição em função de grandezas de entrada, ditas secundárias, de fácil medição. Trata-se de uma alternativa que surge de uma dificuldade operacional ou do alto custo na obtenção da variável desejada (MORAIS JR, 2011).

Desse modo, é notório que os sensores virtuais representam uma estratégia eficaz para otimizar custos e melhorar a eficiência industrial. Enquanto sensores físicos exigem planos de manutenção, calibração e ajuste, a substituição por sensores virtuais pode resultar em economia significativa nas operações fabris. Além disso, sensores virtuais evitam a aquisição desnecessária de sensores físicos, especialmente à medida que a quantidade de sensores em uma planta aumenta. Lidar com grandes volumes de dados provenientes desses sensores é um desafio, mas os sensores virtuais são uma solução prática, permitindo a obtenção de variáveis de interesse a partir de informações já disponíveis de outros sensores, com base em relações matemáticas. Desta forma, a instrumentação virtual surge como uma alternativa viável para qualquer variável que possa ser inferida utilizando métodos matemáticos para monitoramento em tempo real (JÚNIOR, 2015). Em (KADLEC; GABRYS, 2023) o autor disserta acerca de uma metodologia para criação de um sensor virtual, que possui os seguintes 4 principais passos: (1) inspeção inicial dos dados;(2) identificação de estados estacionários;(3) pré-processamento dos dados;(4) seleção do modelo;(5) validação do modelo. O processo é ilustrado na Figura 4:

Figura 4 – Metodologia do *soft sensor*.



Fonte: Autoria Própria.

2.7 Conceito TEDA

O *framework* denominado TEDA (*Typicality and Eccentricity Data Analytics*), que se baseia nos conceitos de tipicidade e excentricidade, foi proposto e desenvolvido por (ANGELOV, 2014). Em essência, o TEDA utiliza esses conceitos para identificar *outliers* em conjuntos de dados. A tipicidade está relacionada à medida de quão semelhante uma amostra de dados é em relação às demais amostras do mesmo conjunto (SILVA, 2022). Por outro lado, a excentricidade avalia o grau de excepcionalidade de um dado em comparação com os demais.

É importante destacar que o TEDA se destaca no campo de *Machine Learning* devido à sua abordagem singular, que segundo (ANGELOV, 2014) elimina a necessidade de:

- Fazer suposições prévias sobre a distribuição dos dados, o que é uma característica comum em muitos métodos de detecção de *outliers*;
- Especificar previamente parâmetros dependentes do problema, o que pode ser uma tarefa desafiadora em algumas situações;
- Pressupor independência entre as amostras de dados individuais, o que é uma restrição em muitos algoritmos tradicionais;
- Requerer um número infinito de observações; de fato, o TEDA demonstra sua eficácia com apenas três amostras de dados, o que é notavelmente eficiente.

No contexto do TEDA, emergem três conceitos cruciais: tipicidade (Γ), excentricidade (ξ) e proximidade acumulada (π). Estes conceitos fornecem uma base fundamental para a análise de dados dentro do TEDA.

Considerando um espaço de dados n-dimensional, representado como $X \in \mathbb{R}^n$, onde a distância entre pontos x e y pode ser definida de várias maneiras, como a distância Euclidiana, as amostras de dados são organizadas como um conjunto de vetores ordenados:

$$X = \{X_1, X_2, \ldots, X_k, \ldots\},\ X_k \in \mathbb{R}^n,\ k \in \mathbb{N}$$

Cada amostra X_k representa o sistema em um momento específico k, permitindo que essa metodologia seja adaptada a conjuntos de dados de diferentes dimensionalidades n (SILVA, 2022). Isso nos permite calcular distâncias para fluxos de dados em diferentes instantes, sendo a primeira $k = 1$ e segunda $k = 2$, para qualquer amostra igual ou superior a 2 é possível realizar o cálculo da distância.

- Proximidade acumulada(π) de um certo ponto x é definida conforme a Equação (2.1) (ANGELOV, 2014).

$$\pi_k(x) = \sum_{i=1}^{k} d(x, x_i), k \geq 2 \tag{2.1}$$

- A excentricidade de uma amostra em um determinado momento k é calculada pela razão entre a sua proximidade acumulada e a soma das proximidades acumuladas de todas as outras amostras, conforme ilustrado na Equação (2.2) do trabalho de (ANGELOV, 2014).

$$\xi_k(x) = \frac{2\pi_k(x)}{\sum_{i=1}^{k} \pi_k(x_i)}, k \geq 2, \sum_{i=1}^{k} \pi_k(x_i) > 0 \tag{2.2}$$

- A tipicidade (Γ) é o complemento da excentricidade, definida por:

$$\Gamma_k(x) = 1 - \xi_k(x) \tag{2.3}$$

Contudo, vale ressaltar que, a tipicidade (Γ) e a excentricidade (ξ) são determinadas com base em um conjunto mínimo de três amostras distintas ($k \geq 3$), uma vez que, de acordo com (ANGELOV, 2014), qualquer par de amostras não idênticas apresenta igual nível de excentricidade e tipicidade.

Armazenar amostras de dados torna-se custoso e há limitações na maioria dos dispositivos. Sendo assim, a excentricidade e tipicidade podem ainda ser calculadas de forma recursiva, desse modo, não é necessário manter um banco de dados registrando dados passados, sendo utilizado apenas a última amostra recebida e os valores que representem o sistema no momento anterior. A forma recursiva é representada pela equação:

$$\xi_k(x) = \frac{1}{k} + \frac{(\mu_k - x_k)^T(\mu_k - x_k)}{k\sigma_k^2} \tag{2.4}$$

Em que a média (μ) e a variância (σ^2) representadas pelas seguintes equações:

$$\mu_k(x) = \frac{k-1}{k}\mu_{k-1} + \frac{1}{k}x_k, \quad \mu_1 = x_1 \tag{2.5}$$

$$\sigma_k^2(x) = \frac{k-1}{k}\sigma_{k-1}^2 + \frac{1}{k-1}\|x_k - \mu_k\|^2, \quad \sigma_1^2 = 0 \tag{2.6}$$

2.7.1 Detecção de *Outliers* com TEDA

A Desigualdade de Excentricidade, derivada do TEDA, estabelece um limiar para a detecção de *outliers*, levando em consideração o conceito estatístico da desigualdade de Chebyshev, a qual certifica que não haverá mais do que $\frac{1}{n^2}$ amostras que excederão uma distância de $m\sigma$ em relação à média (BEZERRA, 2017). Contudo, a Desigualdade de Excentricidade oferece resultados semelhantes sem fazer suposições rígidas sobre a distribuição ou independência dos dados. Assim, tal desigualdade é definida pela equação *(COSTA et al., 2015)*:

$$\zeta_k(x) > \frac{n^2 + 1}{2k} \tag{2.7}$$

Em que $\zeta_k(x)$ representa a excentricidade normalizada dos dados e é calculada pela equação a seguir:

$$\zeta_k(x) = \frac{\xi_k(x)}{2}, \quad \sum_{i=1}^{k} \zeta_i(x) = 1, \quad k \geq 2 \tag{2.8}$$

Sendo n um parâmetro que indica a sensibilidade utilizada no *threshold*, ou seja, quanto maior o valor de n menor a sensibilidade do método e menor a quantidade de outliers detectado. O pseudocódigo do TEDA encontra-se desenvolvido no algoritmo 1:

Algoritmo 1 TEDA: IDENTIFICAÇÃO DE OUTLIER.

Enquanto X estiver ativo **Faça**
  Ler a amostra $x_k \in X$
  Se $k == 1$ **Então**
    $\mu_k = x_k$
    $\sigma_k^2(x_k) = 0$
  Senão
    atualizar μ_k usando equação (2.5);
    atualizar $\sigma_k^2(x_k)$ usando equação (2.6)
    atualizar $\xi_k(x)$ usando equação (2.4)
    atualizar $\zeta_k(x)$ usando equação (2.8)
    Se $\zeta_k(x) > \frac{m^2+1}{2k}$ **Então**
      outlier = verdadeiro
    Senão
      outlier = falso
    Fim Se
  Fim Se
  $k = k + 1$
Fim Enquanto

2.8 Métricas de avaliação dos modelos

Para determinar a qualidade e o desempenho dos modelos desenvolvidos, é fundamental realizar uma avaliação rigorosa. Isso envolve a quantificação do aprendizado alcançado durante o treinamento e a comparação das previsões do modelo com dados de teste independentes, reservados para essa finalidade. Desse modo, nesse trabalho de conclusão de curso as métricas utilizadas no modelo estão representadas abaixo e, todas amplamente utilizados para avaliar a precisão dos modelos de séries temporais.

Mean Squared Error (MSE): O MSE, que em português significa Erro Quadrático Médio, calcula a média dos quadrados dos erros para um número N de amostras entre as previsões do modelo ($\overline{y_i}$) e os valores reais (y_i). Ele fornece uma medida quantitativa da dispersão dos erros, destacando valores discrepantes e sua magnitude.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2 \tag{2.9}$$

Mean Absolute Percentage Error (MAPE): O MAPE, que em português significa Erro Absoluto Percentual Médio, avalia o erro médio absoluto em relação aos valores reais, expresso como uma porcentagem. Isso permite entender o desempenho relativo do modelo

em diferentes partes do conjunto de dados.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \overline{y_i}|}{|y_i|} \cdot 100\% \tag{2.10}$$

Mean Absolute Error (MAE): O MAE, que em português significa Erro Absoluto Médio, representa a média dos valores absolutos dos erros entre as previsões e os valores reais. Ele oferece uma visão direta do tamanho médio dos erros.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \overline{y_i}| \tag{2.11}$$

Root Mean Squared Error (RMSE): O RMSE, que em português significa Raiz do Erro Quadrático Médio, é a raiz quadrada do MSE e é amplamente utilizado para avaliar a precisão geral do modelo, considerando tanto a magnitude quanto a dispersão dos erros.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2} \tag{2.12}$$

Essas métricas fornecem *insights* valiosos sobre o desempenho do modelo, permitindo uma análise completa de sua capacidade de fazer previsões precisas. Idealmente, deseja-se minimizar todas essas métricas, indicando que o modelo é capaz de fornecer previsões próximas aos valores reais.

2.9 Análise de Incertezas baseados no *Dropout*

De acordo com o (INMETRO, 2008) a incerteza associada a uma certa medição caracteriza a dispersão dos valores de um certo mensurando, podendo ter como parâmetro um desvio-padrão em prol de estabelecer um intervalo específico de confiança para tal variável. Assim, todo instrumento de medição apresenta um erro associado a sua medição, de modo que a medição real da variável deve estar contido no intervalo de confiança, que pode ser descrito como o valor medido mais ou menos o desvio padrão, conforme a Equação (2.13). Aqui, i representa o intervalo de confiança da medição, v é o valor medido e V_m é o valor verdadeiro da variável medida (CAMPILHO, 2011).

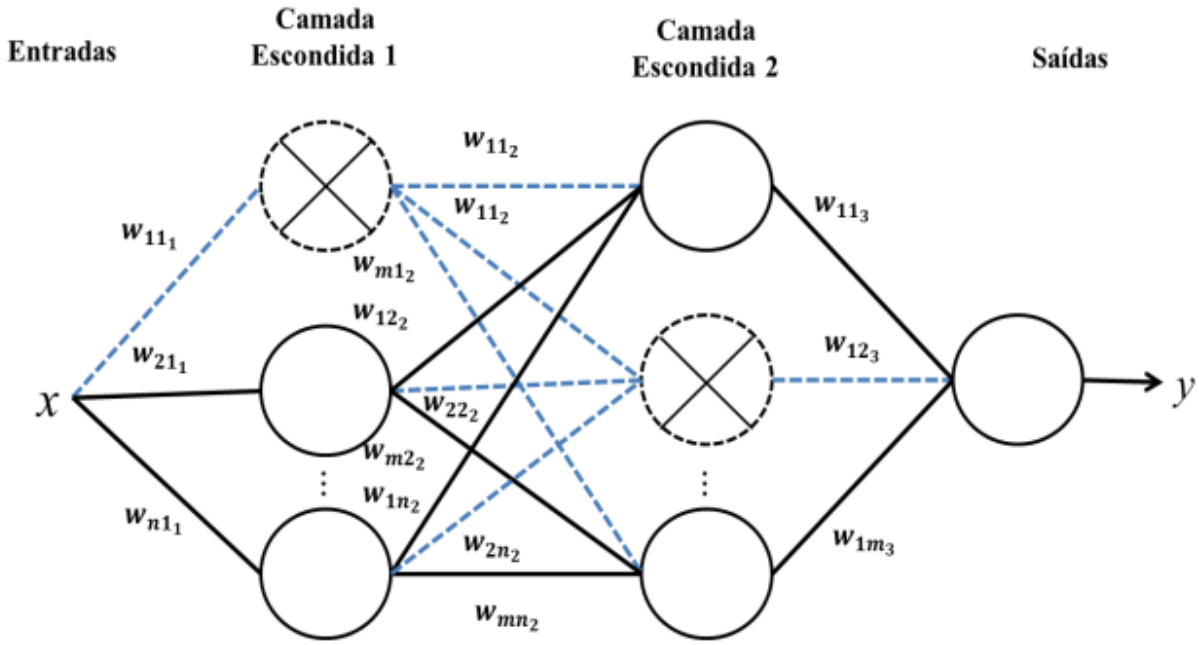
$$i = [v - \sigma, v + \sigma]\ V_m \in i \tag{2.13}$$

2.9.1 Avaliação da incerteza em RNAs por meio de Monte Carlo Dropout

No desenvolvimento de sensores virtuais baseados em Redes Neurais Artificiais uma abordagem utilizada para medição de sua incerteza é o Monte Carlo Dropout. O Monte

Carlo Dropout é uma abordagem que se baseia no conceito de *dropout*, uma técnica de regularização utilizada em Redes Neurais Artificiais para evitar o $overfitting$, tornando o modelo mais geral. O método consiste em desativar, de forma aleatória, neurônios nas camadas de entrada e camadas ocultas da RNA, conforme a Figura 5.

Figura 5 – Arquitetura da Rede Neural com desativação de neurônios.



Fonte: (LIMA, 2022a).

Na técnica usual de regularização *dropout* os neurônios são desativados apenas na fase de treinamento da rede, realizando ajustes nos resultados e todos os neurônios são ativados na fase de teste. Entretanto, a técnica de Monte Carlo Dropout propõe que o *dropout* continue ativo durante a fase de teste, permitindo que sejam geradas T previsões diferentes para uma mesma entrada, levando em consideração a aleatoriedade do processo de *dropout*. A dispersão dessas previsões pode ser usada como uma medida da incerteza inerente ao modelo, conhecida como incerteza epistêmica. Em outras palavras, o Monte Carlo Dropout fornece uma estimativa da incerteza associada às previsões do sensor virtual, possibilitando calcular uma média e uma variância e, a partir disso, calcular o intervalo de confiança para as medições, o que é fundamental em aplicações que requerem alta confiabilidade, como em sistemas de monitoramento e controle em tempo real. Tais parâmetros são calculados pelas seguintes Equações (2.14) e (2.15) (GAL, 2015).

$$\mathbb{E}(\hat{y}) = \frac{1}{N}\sum_{t=1}^{N}\hat{y}_t \tag{2.14}$$

$$\text{Var}(\hat{y}) \approx \tau^{-1} + \frac{1}{N}\sum_{t=1}^{N} \hat{y}_t^N \hat{y}_t - \mathbb{E}(\hat{y})^N \mathbb{E}(\hat{y}) \tag{2.15}$$

Sendo assim, a precisão do modelo (τ) é determinada por meio de certos hiperparâmetros como o peso de regularização L2 (λ), a probabilidade de retenção (1-p) e *length scale* (l) que refere-se a confiança na entrada dos dados. A Equação (2.16) representa a precisão (GAL, 2015):

$$\tau = \frac{l^2 \cdot (1-p)}{2 \cdot N \cdot \lambda} \tag{2.16}$$

Desse modo, o Monte Carlo possui os parâmetros dispostos na tabela 1 relacionados ao seu algoritmo que podem ser variados no desenvolvimento do modelo para obtenção de uma melhor performance.

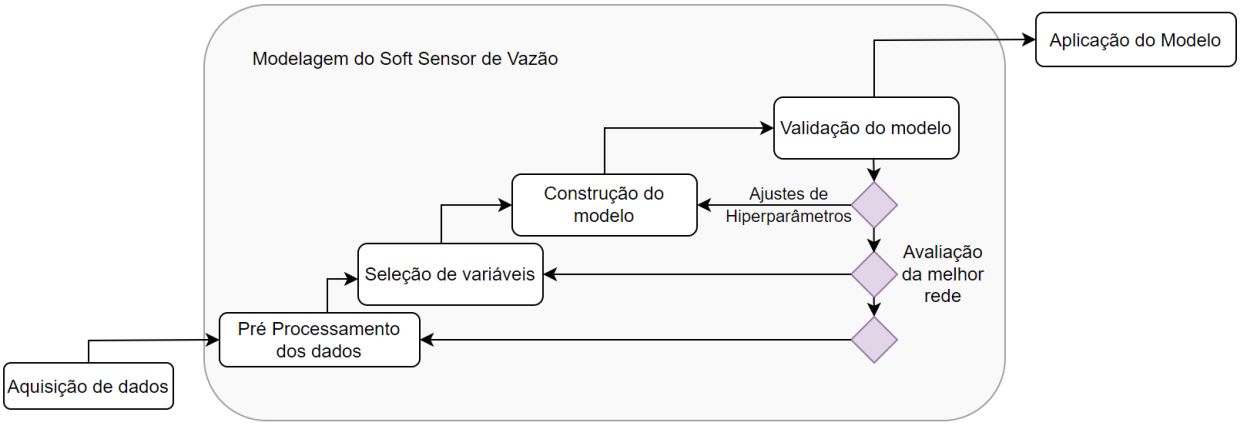
Tabela 1 – Parâmetros do Monte Carlo dropout.

Parâmetro	**Definição**
p	Probabilidade de um neurônio aleatório de uma camada ser desativado
l	Length Scale
N	Número de amostras de treinamento
λ	Peso da regularização L2
T	Número de previsões

3 Metodologia

Nesta seção será apresentada a metodologia para modelagem do *soft sensor* de vazão baseado em técnicas de inteligência artificial. Essa metodologia realiza o uso extensivo de dados associados a grandezas hidráulicas da CAGEPA, com a finalidade de proporcionar um algoritmo computacional que estime a vazão e que forneça um intervalo de confiança associado a tal estimação (incerteza). Para alcançar esse objetivo serão descritas 6 etapas conforme está ilustrado na Figura 6.

Figura 6 – Metodologia do processo.



Fonte: Autoria Própria.

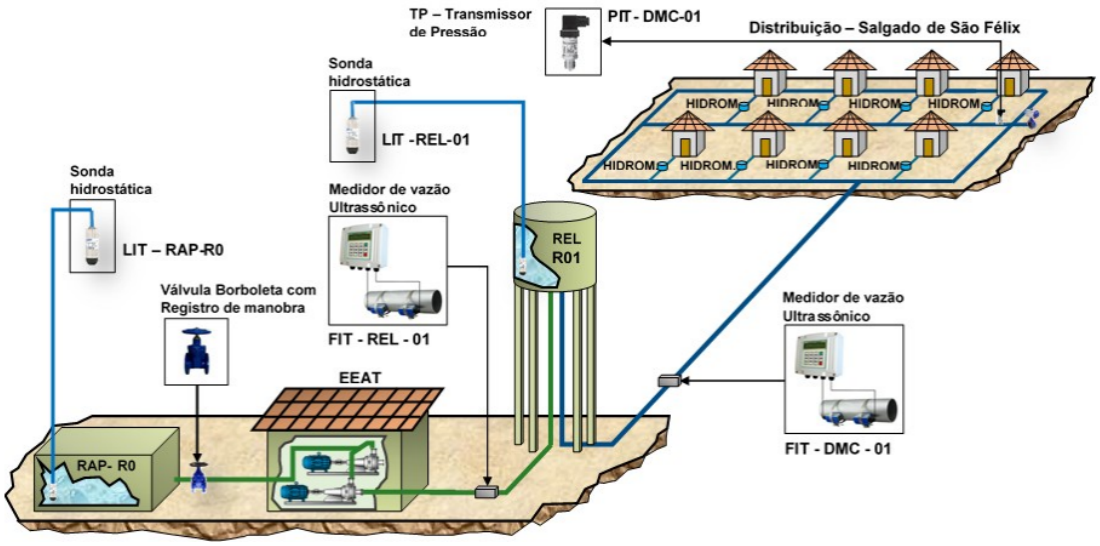
3.1 Aquisição dos dados

Para aquisição do conjunto de dados, foram empregadas informações reais coletadas e transmitidas ao sistema de supervisão em operação pela CAGEPA, na estação de distribuição de água da cidade de Salgado de São Félix. Esse sistema é responsável pela monitorização de um ponto-chave que inclui um reservatório elevado (REL) de distribuição e uma rede de distribuição que atende a aproximadamente 1,4 mil conexões, todas elas sujeitas a monitoramento de pressão. Todos os dados foram tratados utilizando a linguagem de programação Python e suas bibliotecas no ambiente do Google Colab.

3.1.1 Rede de abastecimento de água

O esquema de abastecimento de água que está ilustrado na Figura 7 representa o sistema da CAGEPA na cidade de Salgado de São Félix.

Figura 7 – SAA Salgado de São Felix.



Fonte: *(RAPOSO et al., 2023).*

Os registros referentes a nível, pressão e vazão são encaminhados para o sistema SCADA (*Supervisory Control And Data Acquisition*) da CAGEPA. Tais medições são realizadas a cada 30 segundos e o registro feito a cada 6 minutos representa a média das leituras no intervalo de tempo.

3.1.2 Conjunto de Dados

O DataFrame de dados é composto por cinco variáveis, sendo quatro variáveis secundárias e a vazão (FIT-DMC-01) como sendo nossa variável de interesse, contendo 14,320 amostras de cada. Por meio de estudos com profissionais da área foi decidido permanecer com todas as variáveis disponíveis para o processo. As variáveis utilizadas foram:

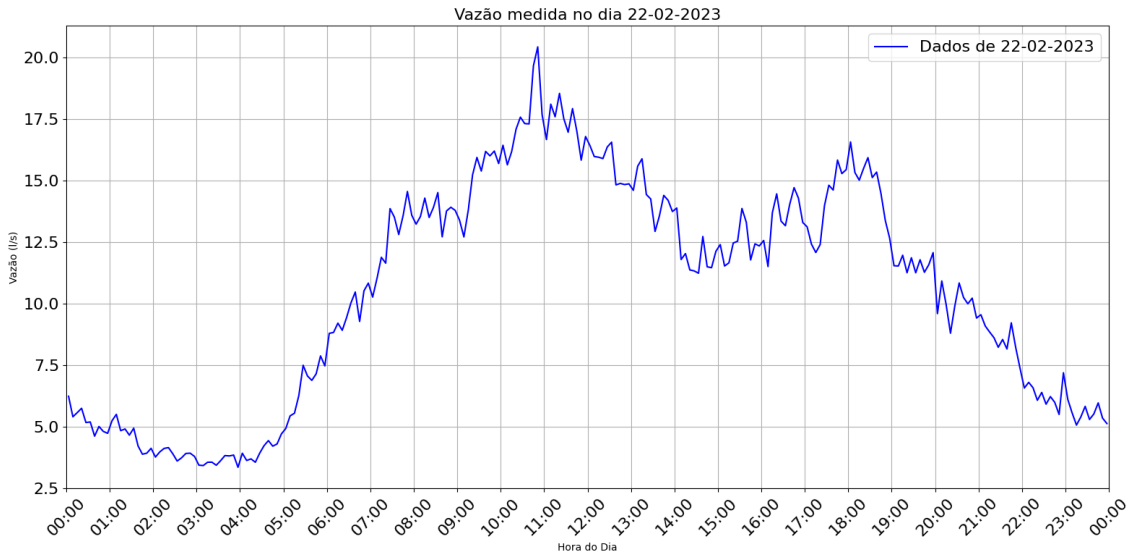
- **L**IT-REL-01 – nível do reservatório de distribuição (m);
- **P**IT-DMC-01 – pressão da rede de distribuição (mca);
- **F**IT-DMC-01 – vazão de saída do reservatório de distribuição (l/s);
- **F**IT-REL-01 - vazão recalque (l/s);
- **P**ressão recalque (mca);

3.1.3 Análise inicial dos dados

Os dados foram colhidos entre os meses de janeiro de 2023 e março de 2023, com uma granularidade de 6 minutos e sendo disponibilizados em formato csv. Na figura 9,

encontra-se ilustrado um gráfico gerado no python utilizando a biblioteca matplotlib que demonstra o comportamento da vazão em um dia no mês de fevereiro.

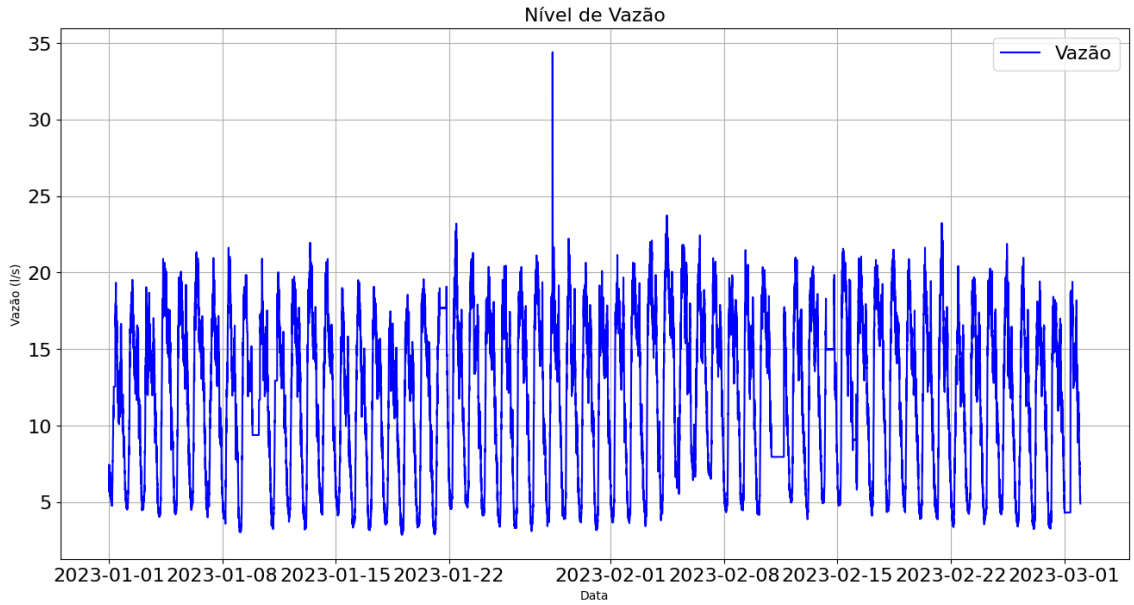
Figura 8 – Dados de vazão colhidos para um dia.



Fonte: Autoria Própria.

Pode-se ainda expandir a visualização para todos os três meses de análise, a Figura 9 representa esse cenário.

Figura 9 – Dados de vazão colhidos para três meses.



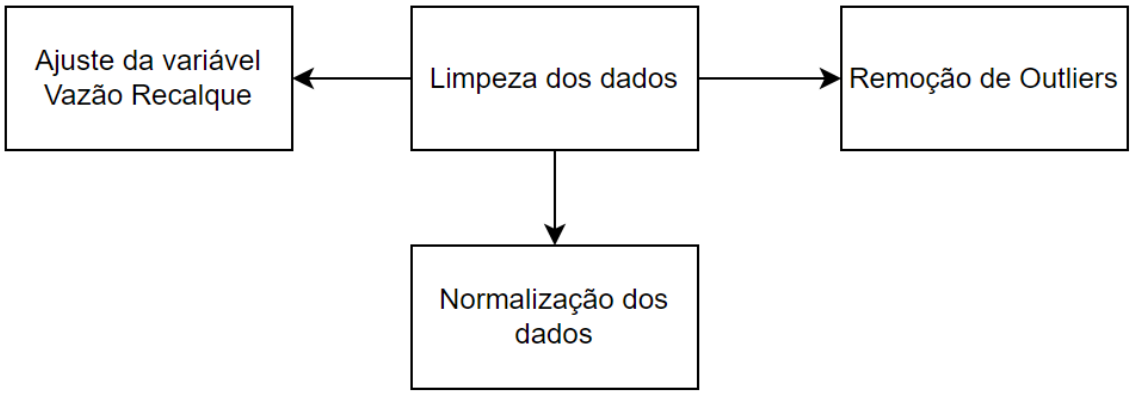
Fonte: Autoria Própria.

Ao realizar a análise da Figura 9 percebemos certos dados com um comportamento diferente do usual, dando a indicação de *outliers* em nossas amostras, sendo assim, era necessário realizar o tratamento desses dados.

3.2 Tratamento dos dados

Antes de realizar o treinamento dos modelos, é necessário fazer o tratamento dos dados, visando remover ruídos e *outliers*. Dessa maneira, a metodologia para o pré-processamento de dados implementada nesse trabalho de conclusão de curso pode ser resumida pelas três principais etapas demonstradas na figura 10:

Figura 10 – Pré Processamento de Dados.



Fonte: Autoria própria.

3.2.1 Tratamento da variável Vazão Recalque

Em um primeiro momento, ao realizarmos a análise dos dados foi evidenciado que a variável de vazão recalque era uma representatividade da ativação da bomba de água para aumentar o nível do reservatório. Sendo assim, tais valores que estavam abaixo de zero representavam o momento em que a bomba estava desligada e acima de zero quando ocorria sua ativação, então, seus dados foram atualizados para uma variável booleana, sendo zero quando estivesse desligado e um quando fosse ligado.

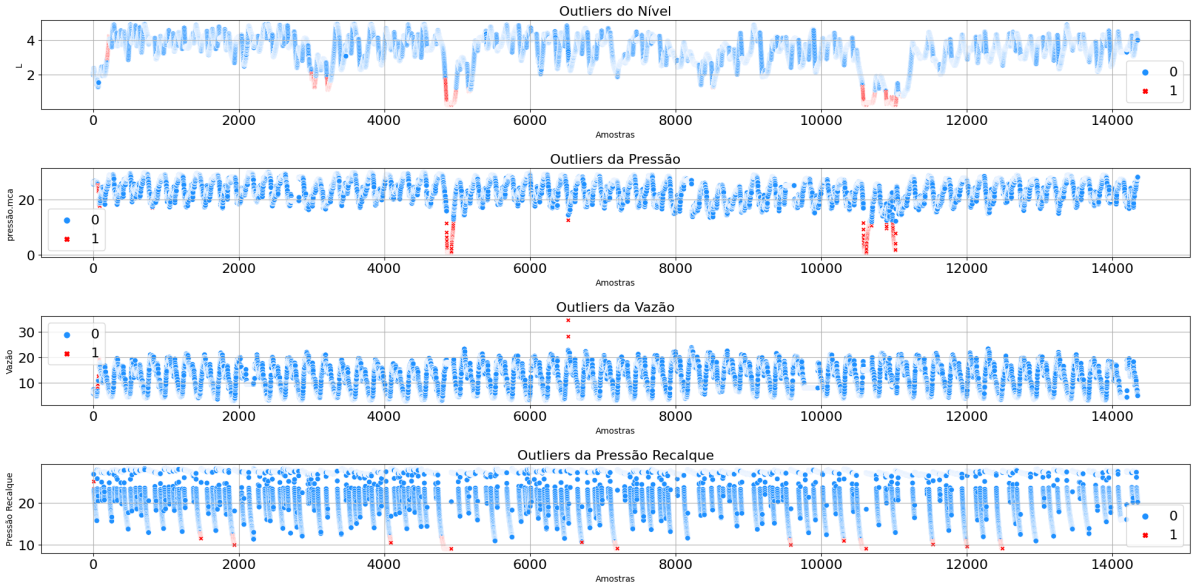
3.2.2 Remoção de *outliers*

Uma das formas principais de identificação de *outliers* acontece por meio do limiar $n\sigma$ (BERNIERI; BETTA; LIGUORI, 1996), em que há o cálculo dos dados anormais que não estejam no intervalo $[\overline{y} - 3\sigma, \overline{y} + 3\sigma]$, sendo $\overline{y}$ a média do conjunto de dados e σ o desvio padrão. Outra proposta, a qual é utilizada nesse TCC é o *framework* TEDA, calculando o quão excêntrico é um dado dentro de seu conjunto. Para tal utilizamos a seguinte fórmula da desigualdade de excentricidade :

$$\zeta_k > \frac{n^2 + 1}{2k}, n = 2.5 \tag{3.1}$$

A escolha do valor $n = 2.5$ foi escolhida a partir da verificação de outros trabalhos acadêmicos por apresentar resultados satisfatórios. Na Figura 11, os dados marcados em vermelho representam os *outliers* detectados através da aplicação do algoritmo do TEDA.

Figura 11 – Identificação de *Outliers*.



Fonte: Autoria própria.

Tais dados apresentados como *outliers* foram substituídos pela média geral naquela hora.

3.2.3 Normalização dos dados

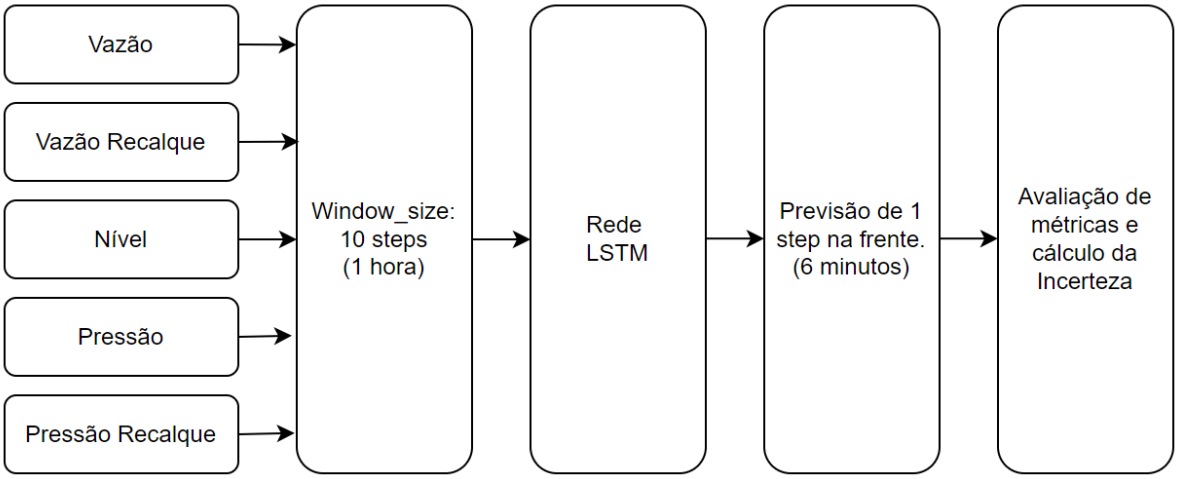
Outra etapa na preparação dos dados para o modelo LSTM foi a de normalização. O objetivo é de escalar os valores de todas as variáveis para o intervalo entre 0 e 1. Essa normalização foi aplicada a todo o conjunto de dados, garantindo que as informações fossem tratadas de forma consistente e facilitando o treinamento do modelo. Essa abordagem é essencial para o desempenho e a convergência eficaz do LSTM, uma vez que reduz as disparidades nas escalas das variáveis, tornando os dados mais adequados para o processo de aprendizado.

3.3 Topologia do Modelo

O modelo desenvolvido para predição da vazão é baseado em Redes Neurais do tipo LSTM motivado principalmente pela sua capacidade de reter memórias de curto prazo e ser comumente utilizada em problemas de séries temporais. A previsão de curtíssimo prazo recebe como entrada os valores das cinco variáveis atrasados em 10 medições, sendo

o equivalente a uma hora e retorna na saída da rede uma previsão de um passo a frente da vazão. O fluxograma do processo pode ser visualizado na Figura 12.

Figura 12 – Topologia do Modelo.



Fonte: Autoria própria.

Foram analisados os desempenhos dos modelos formado pelas cinco variáveis preditivas e a metodologia empregada na identificação da melhor configuração de rede para o desenvolvimento do *soft sensor* no contexto deste projeto consistiu em três etapas:

1. **Definição da Quantidade de Neurônios na Primeira Camada Interna:** Inicialmente, procedemos à variação do número de neurônios na camada de entrada. Essa variação foi realizada de N a $5N$, com N representando a quantidade de passos passados definidos na janela temporal (10 passos);

2. **Identificação da Necessidade de Camadas Adicionais:** Após a determinação da melhor quantidade de neurônios na primeira camada, avaliamos se seria vantajoso incorporar camadas internas adicionais à arquitetura da rede neural, variando novamente a segunda camada de N a $5N$;

3. **Variação dos Parâmetros do Monte Carlo Dropout:** Por fim, procedemos à variação dos parâmetros relacionados à técnica de incerteza. Isso incluiu a variação da taxa de *dropout*, variando de 5% a 15%, e a variação do valor do parâmetro λ, considerando valores como [4.733e-7, 4.733e-8, 9.466e-9, 4.734e-9, 4.733e-11]. A escolha desses valores se deu utilizando a Equação (2.16) variando os valores arbitrários da precisão do modelo (τ) em [1,10,50,100,10000].

A comparação entre os modelos foi feita de acordo com as métricas convencionais de avaliação de modelos de redes neurais, já citados anteriormente, para cada conjunto de variáveis selecionado. Também foi feita a avaliação de incertezas do Modelo, que será calculada por meio da técnica Monte Carlo dropout.

4 Resultados

Esta seção apresenta os resultados de cada etapa da metodologia especificada na Seção 3. Inicialmente, são detalhados os diversos experimentos conduzidos, nos quais houve variação de hiperparâmetros da rede neural. Posteriormente, será apresentado o procedimento de avaliação da incerteza de medição, bem como a comparação entre as configurações testadas. O capítulo abordará os detalhes da criação do modelo de *soft sensor* e a avaliação dos resultados obtidos nos experimentos.

No modelo desenvolvido, com cinco variáveis utilizadas, os dados foram divididos em conjuntos de treinamento (70%), validação (10%), e teste (20%), totalizando 10.035 amostras para o conjunto de treinamento, 1.425 amostras para o conjunto de validação, e 2.860 amostras para o conjunto de teste. Todos foram treinados utilizando a mesma função de otimização Adam e com um número de previsões $T = 10$ para determinação da incerteza do modelo e a média das previsões. A escolha dessa quantidade para o número de previsões se deu através de experimentações ao longo do trabalho, percebendo que sua alta variação não influenciaria em melhores resultados para os dados.

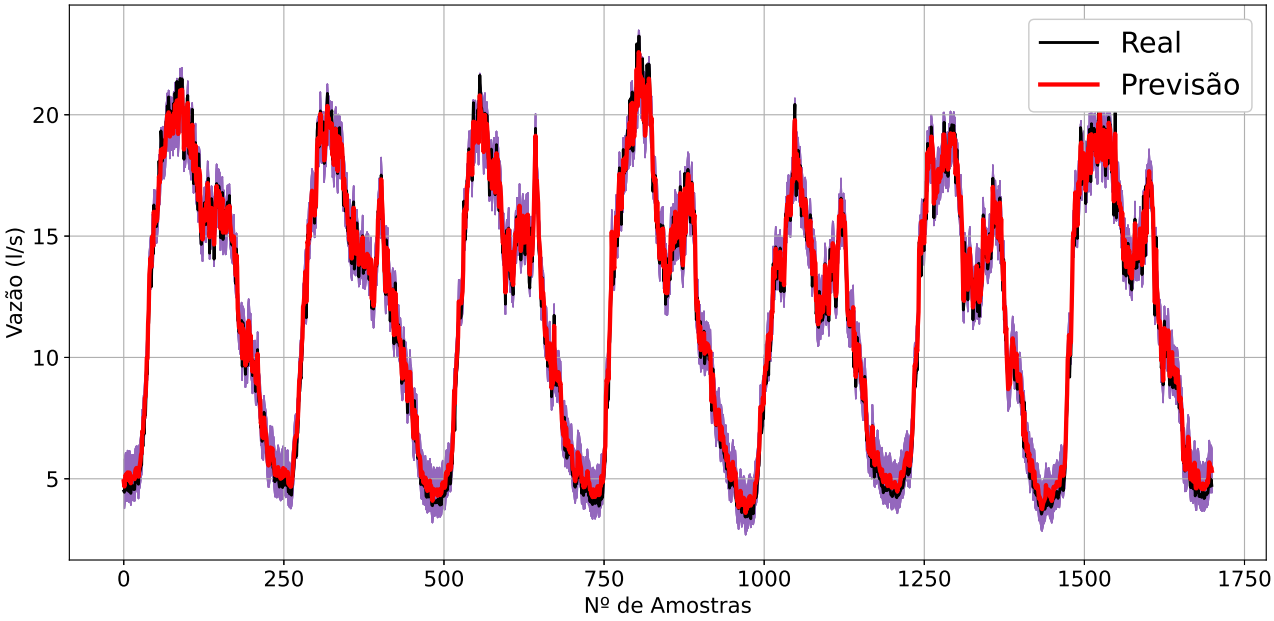
Em um primeiro momento, foi fixado o valor de *dropout*, regularizador do kernel e precisão do modelo (τ), definindo apenas uma camada interna na rede e realizada a variação dos neurônios. com 50 épocas e *batch size* de 32. Na Tabela 2 são apresentados os resultados obtidos com tais variações:

Tabela 2 – Resultados de Variações de Neurônios na Primeira Camada.

Modelos	Neurônios	MSE	RMSE	MAE	MAPE%	Incerteza	Probabilidade Dropout
1	10.0	0.430701	0.656278	0.568334	7.960901	1.805104	5.0
2	20.0	0.122777	0.350396	0.286126	3.552791	1.803993	5.0
3	30.0	0.137360	0.370621	0.310982	3.814102	1.803791	5.0
4	40.0	0.152465	0.390468	0.328037	4.202805	1.803626	5.0
5	50.0	0.128762	0.358834	0.290451	3.702435	1.803533	5.0

Para esse primeiro experimento, observamos que o modelo 2, com 20 neurônios obteve a melhor performance em todas as métricas, com observação ao MAPE com porcentagem de erro baixa. Na Figura 13, estão ilustradas as previsões de uma semana desse modelo, onde a curva na cor preta refere-se aos valores reais naquele instante de tempo e os valores vermelhos as previsões realizadas pelo modelo, o lilás representa a incerteza do modelo.

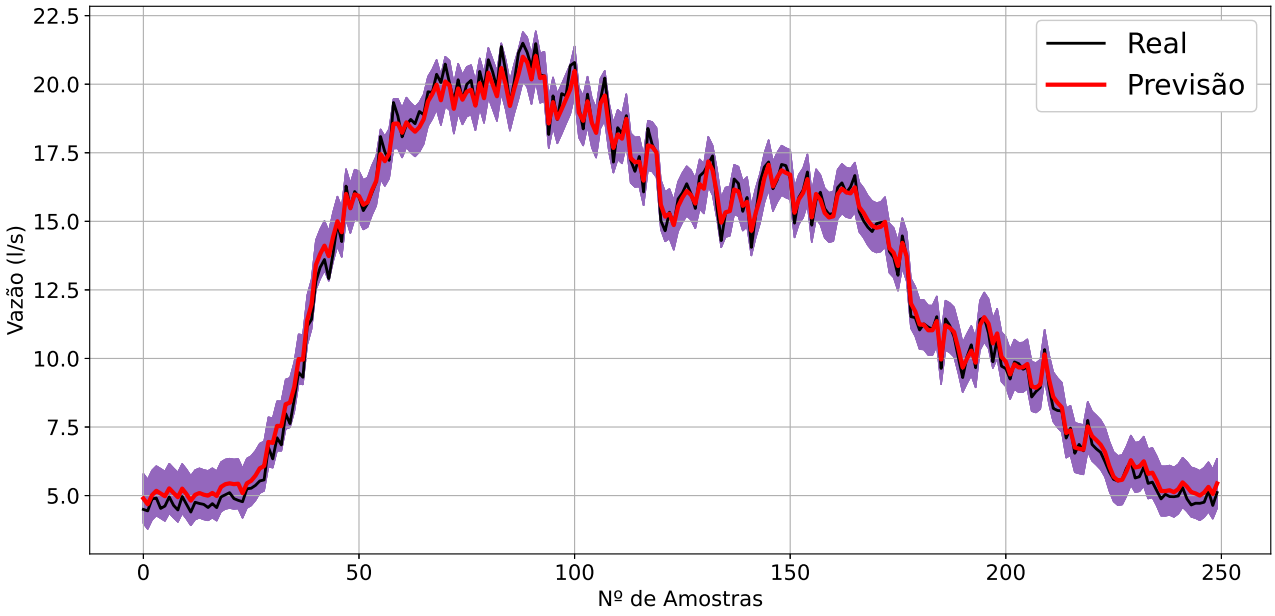
Figura 13 – Previsão do modelo com 1 camada para uma semana.



Fonte: Autoria própria.

Podemos ainda visualizar o gráfico 14 apresentando a previsão de um dia, na intenção de ter uma melhor análise:

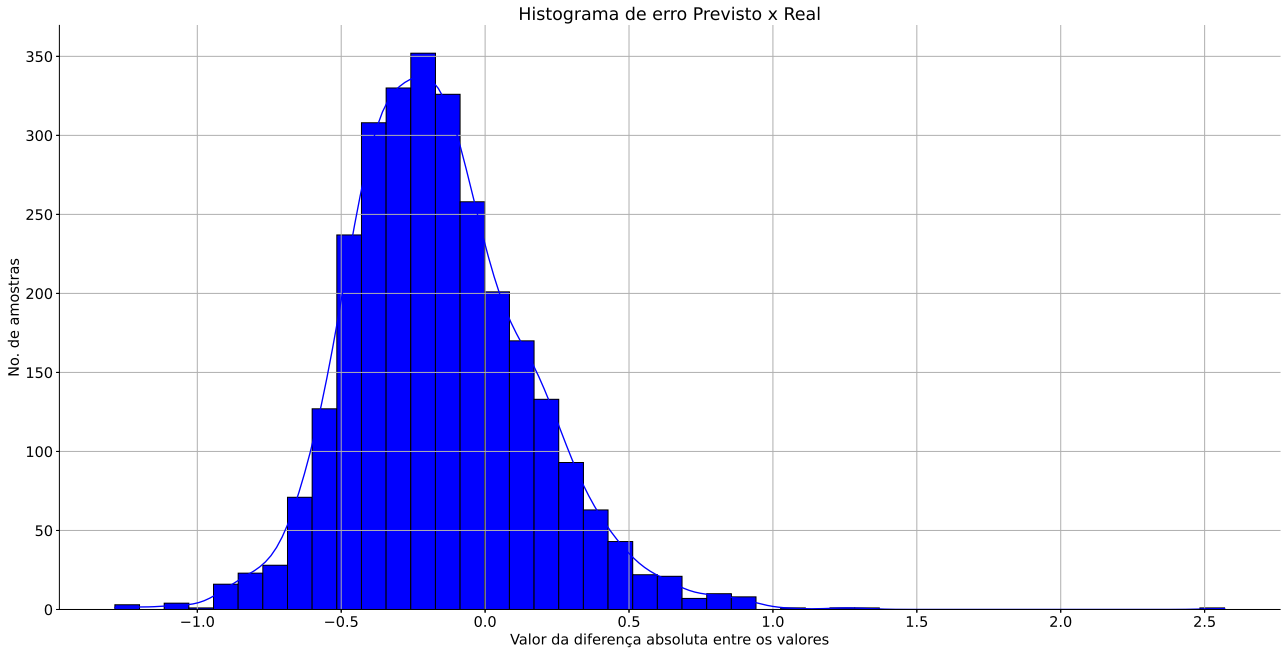
Figura 14 – Previsão do modelo com 1 camada para um dia.



Fonte: Autoria própria.

Ainda nessa perspectiva, a figura 15 demonstra o histograma do erro absoluto para as previsões e os valores reais nesse modelo.

Figura 15 – Histograma de erros absolutos - 1 camada.



Fonte: Autoria própria.

4.1 Influência da Segunda Camada

Uma das considerações fundamentais no aprimoramento de modelos é a exploração de diferentes arquiteturas e configurações. Ao fixar a primeira camada com 20 neurônios passamos a avaliar a influência da adição de uma segunda camada em nosso modelo. Dessa forma, observamos resultados bem similares aos com apenas uma camada, porém a melhor configuração continua sendo com apenas 20 neurônios. Esses resultados são evidenciados pelas métricas, como MAPE e RMSE. A análise dos resultados pode ser visualizada na tabela 3

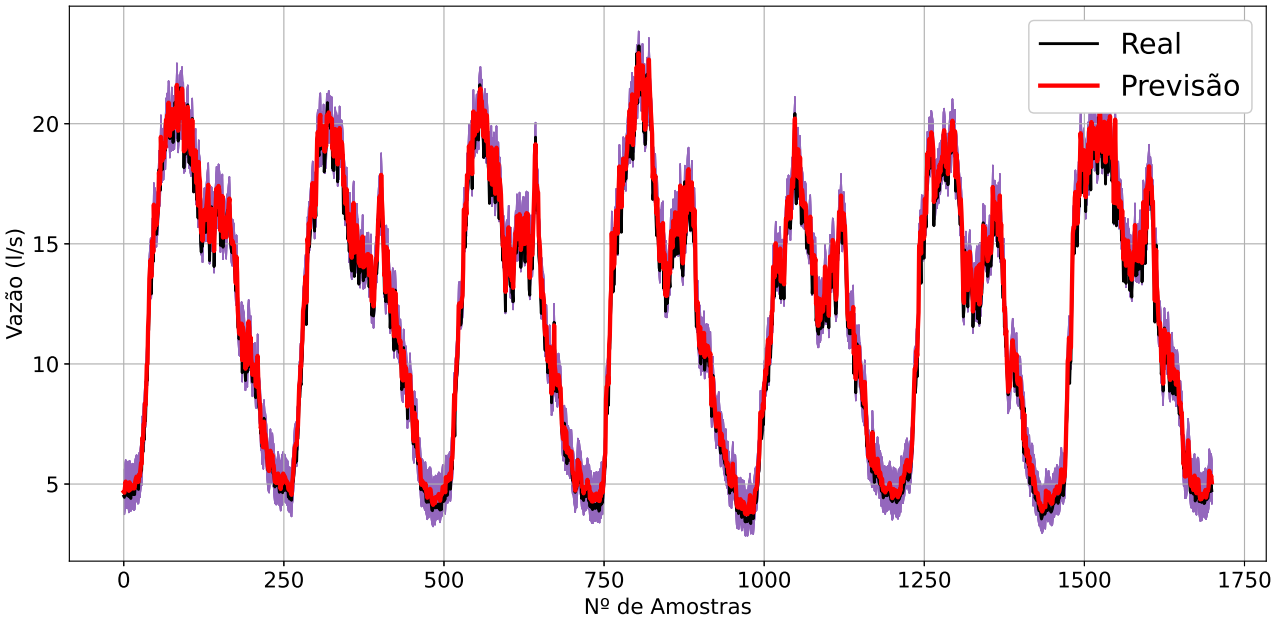
Tabela 3 – Resultados de Variações de Neurônios na Segunda Camada.

Modelos	Neurônios	MSE	RMSE	MAE	MAPE%	Incerteza	Probabilidade Dropout
6	10.0	0.186025	0.431307	0.357726	4.810097	1.804732	5.0
7	20.0	0.201765	0.449183	0.383096	3.973360	1.804100	5.0
8	30.0	0.143061	0.378234	0.318678	4.087328	1.803780	5.0
9	40.0	0.244428	0.494397	0.416760	5.772407	1.803597	5.0
10	50.0	0.138231	0.371795	0.312508	4.084418	1.803490	5.0

Analisando de forma comparativa aos resultados com apenas uma camada, notamos que mesmo sem ser uma diferença significativa, os resultados com apenas uma camada interna performam melhor para todas as métricas avaliadas. As previsões do modelo dois

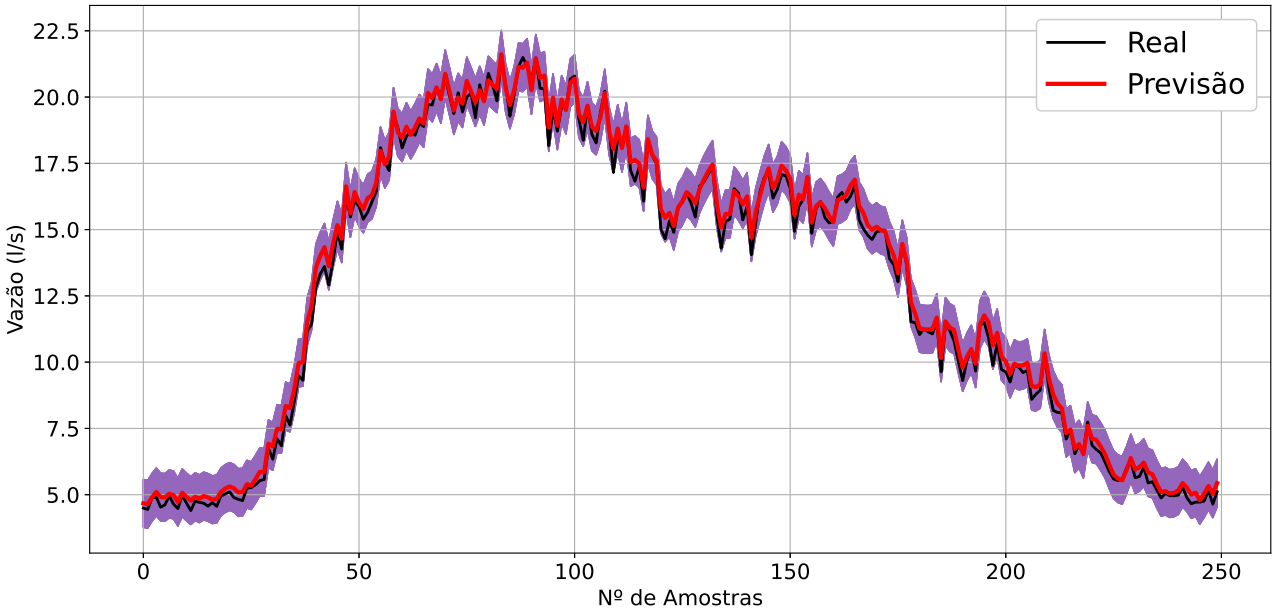
com duas camadas para uma semana e um dia e o histograma do erro absoluto estão dispostos, respectivamente, nas figuras 16, 17 e 18.

Figura 16 – Previsão do modelo para 1 semana com 2 camadas.



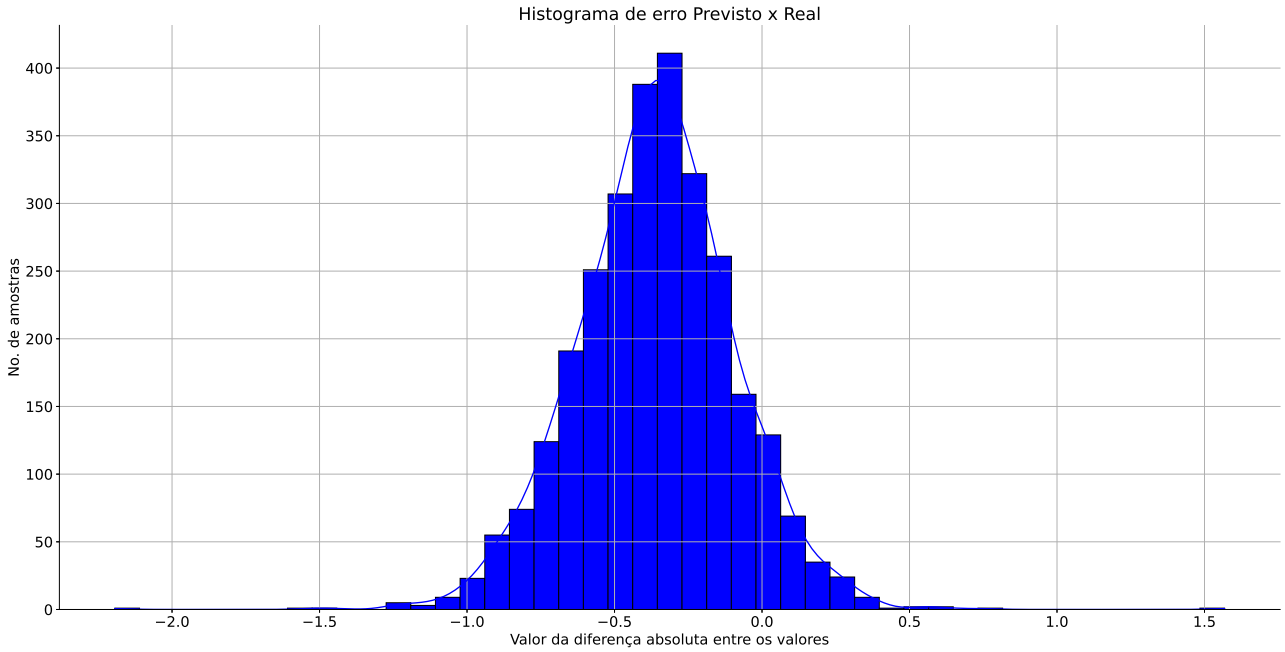
Fonte: Autoria própria.

Figura 17 – Previsão do modelo para 1 dia com 2 camadas.



Fonte: Autoria própria.

Figura 18 – Histograma de erros absolutos - 2 camadas.



Fonte: Autoria própria.

Desse modo, optou-se por realizar a variação dos hiperparâmetros do algoritmo de Monte Carlo Dropout utilizando apenas uma camada com 20 neurônios como foi evidenciado anteriormente.

4.2 Variação de Parâmetros no Monte Carlo

Definida a quantidade de neurônios e de camadas internas, a técnica de Monte Carlo dropout foi utilizada para analisarmos a incerteza do modelo.

Uma espécie de *GridSearch* variando os parâmetros da Tabela 1 foi utilizada conduzindo diversos testes com o intuito de ajustar os hiperparâmetros do modelo de Monte Carlo Dropout baseado em estudos dos trabalhos de (LIMA, 2022a) e (FROES, 2022). Enquanto certos parâmetros, como o número de previsões (T) e o length scale (l), foram mantidos constantes, outros parâmetros, como a probabilidade de dropout (p) e o peso da regularização L2 (λ), foram submetidos a variações sistemáticas. O valor de $l = 0.1$ foi selecionado de forma arbitrária estimando uma incerteza na aquisição de dados de sensores. Com relação ao λ, quanto menor o seu valor, menor a incerteza do modelo. Contudo, não necessariamente o aumento do valor para ter um maior intervalo de confiança trará melhores resultados, uma vez que sua variação impacta nos pesos da Rede Neural, podendo piorar o resultado das previsões. O objetivo dessas variações era encontrar os valores ideais que minimizassem as métricas de desempenho, como o RMSE e o MAPE. Os resultados se encontram na Tabela 4.

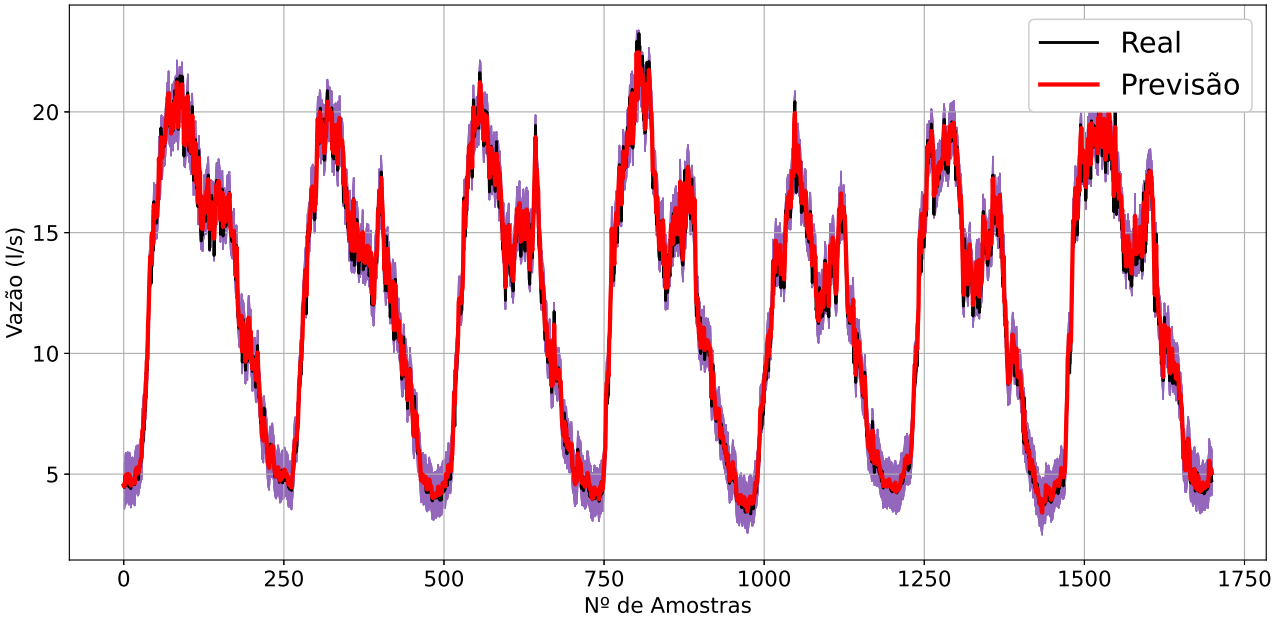
Tabela 4 – Resultados de Variações de Parâmetros Monte Carlo.

Modelo	Probabilidade Dropout %	MSE	RMSE	MAE	MAPE%	Incerteza	λ	τ
11	5.0	0.1905	0.4365	0.3758	4.6533	2.0603	4.733e-7	1
12	5.0	0.0974	0.3121	0.2450	2.5368	1.8290	4.733e-8	10
13	5.0	0.2896	0.5382	0.4786	5.3481	1.8070	9.466e-9	50
14	5.0	0.1797	0.4239	0.3592	4.7209	1.8041	4.734e-9	1e2
15	5.0	0.2548	0.5048	0.4403	5.1836	1.8013	4.733e-11	1e4
16	10.0	0.2847	0.5336	0.4528	6.1147	2.0610	4.484e-7	1
17	10.0	0.1794	0.4236	0.3522	4.2931	1.8298	4.484e-8	10
18	10.0	0.1880	0.4336	0.3518	4.8005	1.8077	8.968e-9	50
19	10.0	0.1855	0.4307	0.3610	4.5060	1.8051	4.484e-9	1e2
20	10.0	0.1800	0.4243	0.3487	4.6788	1.8021	4.484e-11	1e4
21	15.0	0.2288	0.4783	0.3901	5.2112	2.0620	4.235e-7	1
22	15.0	0.2975	0.5454	0.4759	5.8930	1.8307	4.235e-8	10
23	15.0	0.4890	0.6993	0.5743	8.5758	1.8085	8.470e-9	50
24	15.0	0.1366	0.3697	0.2918	3.4911	1.8060	4.235e-9	1e2
25	15.0	0.1437	0.3790	0.3062	3.5855	1.8034	4.235e-11	1e4

Ao ajustar e otimizar os parâmetros, conforme detalhado na Tabela 4, destaca-se o teste identificado como sequência número 12, o qual alcançou um mínimo valor em todas as métricas calculadas. Nesse contexto, a incerteza de medição do modelo se situa em torno de 1,8 vezes o desvio padrão, um indicador altamente relevante na previsão de vazão de água para um Sistema de Abastecimento de Água. Essa precisão se revela particularmente significativa, garantindo que as demandas do SAA sejam atendidas de maneira eficaz e confiável.

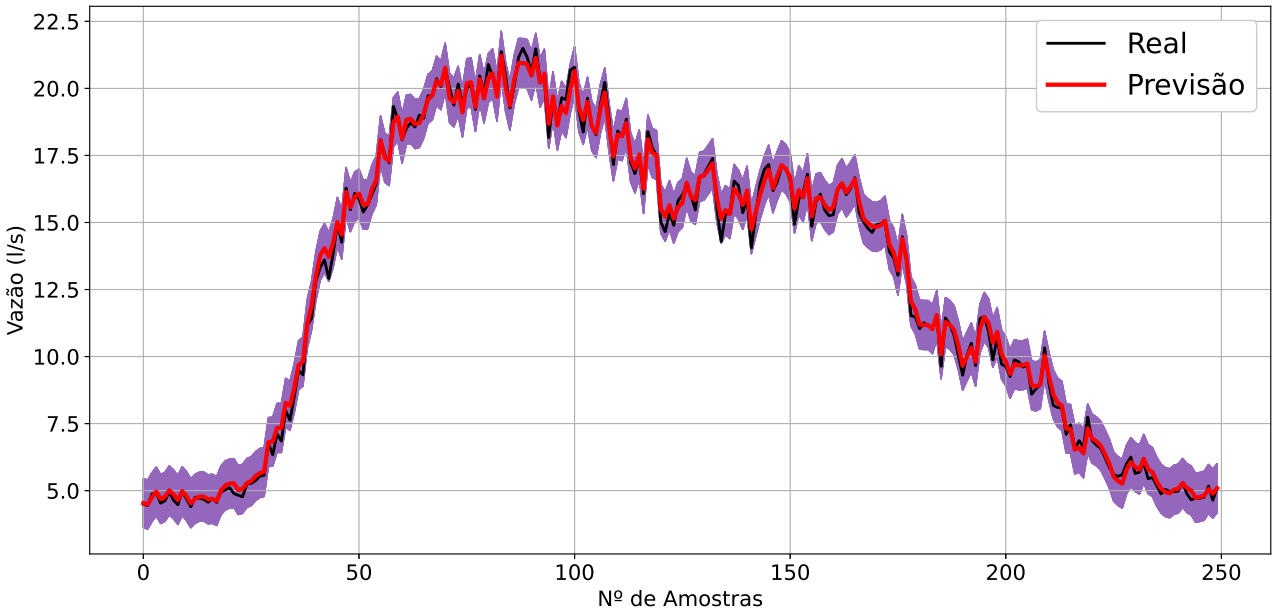
Os resultados podem ser visualizados na representação gráfica a seguir, em que a curva em preto corresponde à vazão de água real, enquanto a curva vermelha representa a vazão prevista pelo modelo proposto. A faixa sombreada em lilás, delimitada pelas duas curvas, representa a incerteza de medição.

Figura 19 – Previsão do modelo para 1 semana - Melhor Modelo.



Fonte: Autoria própria.

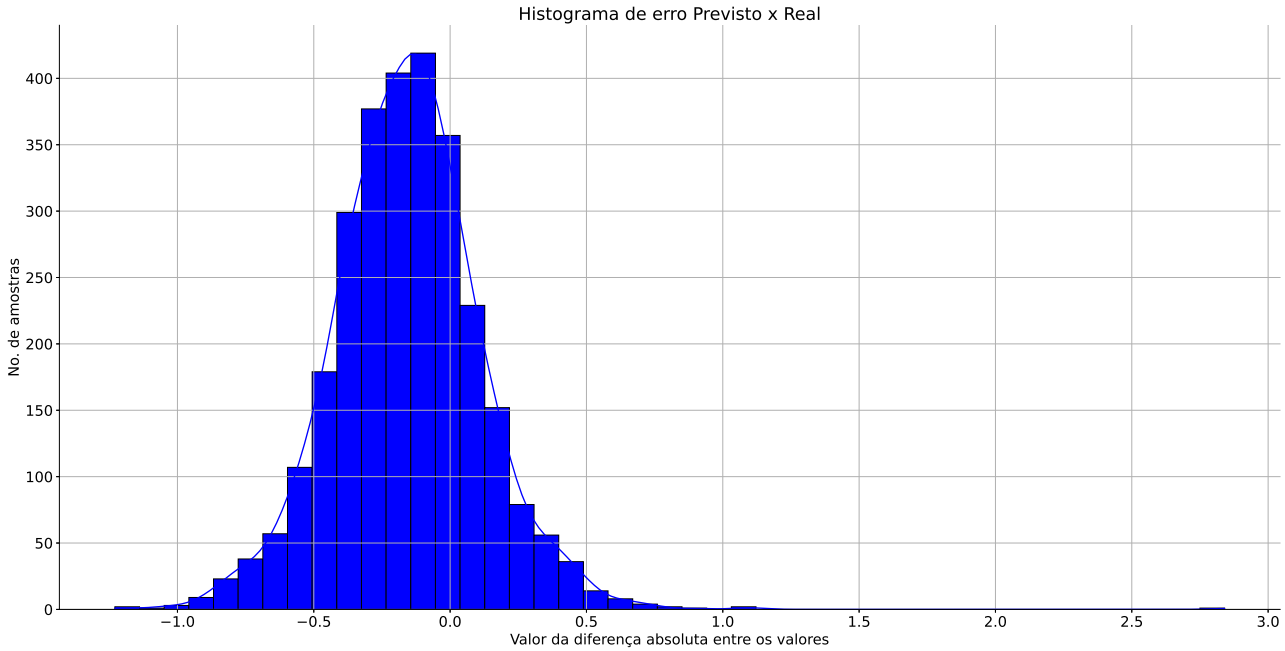
Figura 20 – Previsão do modelo para 1 dia - Melhor Modelo.



Fonte: Autoria própria.

Para identificarmos o comportamento do erro, plotamos o histograma associado ao modelo:

Figura 21 – Histograma de erros absolutos - Melhor Modelo.



Fonte: Autoria própria.

5 Conclusão

Neste estudo, foi realizada a coleta de dados de um sistema de abastecimento de água da cidade de Salgado de São Félix. Realizou-se uma análise inicial dos dados e através do algoritmo TEDA foi possível realizar a correção dos outliers, substituindo tais valores pela média geral naquela hora do dia. Posteriormente foi desenvolvido e avaliado a implementação de um *soft sensor* baseado em Redes Neurais LSTM para a previsão de vazão em sistemas de abastecimento de água.

Os resultados obtidos nas análises dos modelos testados revelam a eficácia das redes neurais recorrentes LSTM, mesmo em configurações de menor complexidade com uma única camada interna. Essas redes demonstraram uma notável capacidade de predição de padrões em amostra de dados de vazão, evidenciada por baixos valores de MAPE e RMSE, como observado no Modelo 2 com um MAPE de 3,55% e um RMSE de 0,65, bem como no Modelo 7 com duas camadas internas com um MAPE de 3,97% e um RMSE de 0,44. Esses resultados indicam que as redes LSTM são eficazes na previsão precisa do comportamento da vazão, mesmo em cenários mais simples.

Levando em conta o melhor resultado obtido nas variações de parâmetros das redes neurais, o modelo 12, possui um MAPE de 2,53% e RMSE de 0,31. A análise de incertezas, usando o método de Monte Carlo Dropout, validou a precisão do modelo proposto. Com um erro de medição próximo a 1,82(l/s), nossa abordagem se mostrou adequada para a medição de vazão de água.

Diante da crescente digitalização dos sistemas e do contexto da Indústria 4.0, o uso de soft sensors baseados em redes neurais LSTM oferece uma alternativa eficaz para a medição e previsão de vazão de água. Essa flexibilidade e capacidade de reconhecimento de padrões são vantagens significativas, permitindo a substituição ou o complemento de sensores físicos tradicionais.

Este estudo ressalta a importância contínua de pesquisar e aprimorar modelos de SoftSensors para atender a diversas necessidades de monitoramento e controle em sistemas de abastecimento de água.

Para futuras pesquisas, é sugerida a exploração de outras arquiteturas de Redes Neurais, incluindo Redes NARX (*Nonlinear AutoRegressive with eXogenous inputs*), Redes Neurais MLPs e modelos baseados em Árvores de Decisão.

Em resumo, a implementação do soft sensor baseado em redes neurais LSTM oferece uma solução eficiente e econômica para o monitoramento e controle de sistemas de abastecimento de água, contribuindo para a gestão sustentável dos recursos hídricos.

Referências

ALGOLIA. *What Is Explainable AI and Why Is Transparency So Important for Machine Learning Solutions?* 2023. Disponível em: <https://www.algolia.com/blog/ai/what-is-explainable-ai-and-why-is-transparency-so-important-for-machine-learning-solutions/>. Citado na página 18.

ANGELOV, P. Outside the box: An alternative data analytics framework. *Journal of Automation, Mobile Robotics & Intelligent Systems*, v. 8, p. 29–35, 2014. Citado 3 vezes nas páginas 21, 22 e 23.

BARBOSA, G. et al. Segurança em redes 5g: Oportunidades e desafios em detecção de anomalias e predição de tráfego baseadas em aprendizado de máquina. In: _____. [S.l.: s.n.], 2021. p. 145–189. ISBN 9786587003658. Citado na página 20.

BERNIERI, A.; BETTA, G.; LIGUORI, C. On-line fault detection and diagnosis obtained by implementing neural algorithms on a digital signal processor. *IEEE Transactions on Instrumentation and Measurement*, v. 45, n. 5, p. 894–899, 1996. Citado na página 31.

BEZERRA, C. G. *Uma abordagem baseada em tipicidade e excentricidade para agrupamento e classificação de streams de dados*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2017. Citado na página 23.

BRASIL, G. do. *Lei nº 14.026, de 15 de julho de 2020. Estabelece o novo Marco Legal do Saneamento Básico*. 2020. Disponível em: <https://www.gov.br/ana/pt-br/assuntos/saneamento-basico/novo-marco-legal-do-saneamento>. Citado na página 14.

CAMPILHO, A. *Instrumentaçao Electronica: METODOS E TECNICAS DE MEDIÇAO*. [S.l.]: FEUP EDIÇOES, 2011. ISBN 9789727521630. Citado na página 25.

COSTA, B. et al. Online fault detection based on typicality and eccentricity data analytics. In: . [S.l.: s.n.], 2015. Citado na página 23.

FROES, F. *Comparação de modelos de desenvolvimento de sensores virtuais baseados em inteligência artificial*. 2022. Citado na página 38.

GAL, Y. *Probabilistic Deep Learning*. 2015. Disponível em: <https://www.cs.ox.ac.uk/people/yarin.gal/website/blog_3d801aa532c1ce.html>. Citado 2 vezes nas páginas 26 e 27.

INMETRO. *Avaliação de dados de medição — Guia para a expressão de incerteza de medição*. 2008. Disponível em: <https://www.gov.br/ana/pt-br/assuntos/saneamento-basico/novo-marco-legal-do-saneamento>. Citado na página 25.

JÚNIOR, A. A. d. M. *Uso de sensores virtuais (soft sensors) para estimativa de impurezas em colunas de destilação de alta pureza.* Tese (Doutorado) — Universidade Federal de Campina Grande, 2015. Citado na página 21.

KADLEC, P.; GABRYS, B. Application of computational intelligence techniques to process industry problems. 09 2023. Citado na página 21.

LIMA, J. *CONSTRUÇÃO DE SENSOR VIRTUAL PARA MEDIÇÃO DE VAZÃO EM UMA USINA DO SETOR SUCROENERGETICO BASEADO EM REDES NEURAIS ARTIFICIAIS*. Dissertação (Mestrado) — Universidade Federal da Paraíba, 2022. Citado 2 vezes nas páginas 26 e 38.

LIMA, R. *DESENVOLVIMENTO DE UM SOFT SENSOR PARA ESTIMAÇÃO DA VAZÃO EM SISTEMAS DE ABASTECIMENTO DE ÁGUA UTILIZANDO REDES NEURAIS ARTIFICIAIS*. Tese (Doutorado) — Universidade Federal da Paraíba, 2022. Citado na página 14.

MATSUMOTO, F.; DUARTE, G.; MURAKAMI, L. *Redes LSTM*. 2019. Disponível em: <https://medium.com/turing-talks/turing-talks-27-modelos-de-prediÃğÃčo-lstm-df85d87ad210>. Citado na página 19.

PEREIRA, L. L. d. M. *Soft sensor empregando rede neural recorrente LSTM para estimação da gramatura numa máquina de papel*. 2023. Citado na página 19.

RAPOSO, G. et al. Aplicação de soft sensors para macromedição: Um estudo de caso real. In: *32º Congresso Brasileiro de Engenharia Sanitária e Ambiental*. [S.l.: s.n.], 2023. Citado na página 29.

ROCHA, J. d. C. *Predição do Consumo de Água por Meio de Redes Neurais Artificiais – Um Estudo de Caso em Belém-PA*. 2018. Trabalho de Conclusão de Curso, Universidade Federal do Pará. Citado na página 13.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. New Jersey: Pearson Prentice Hall, 2010. Citado na página 17.

SCHLEDER, G. R. et al. From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, IOP Publishing, v. 2, n. 3, p. 032001, may 2019. Disponível em: <https://dx.doi.org/10.1088/2515-7639/ab084b>. Citado na página 18.

SILVA, M. B. D. d. *Uma metodologia orientada a fluxo de dados para modelagem do comportamento de motoristas*. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2022. Citado 2 vezes nas páginas 21 e 22.

SNIS. *Mapa de Indicadores de Água*. 2021. Disponível em: <http://appsnis.mdr.gov.br/indicadores/web/agua_esgoto/mapa-agua>. Citado na página 13.

SOARES, P.; SILVA, J. da. Aplicação de redes neurais artificiais em conjunto com o método vetorial da propagação de feixes na análise de um acoplador direcional baseado em fibra Ótica. *Revista Brasileira de Computação Aplicada*, v. 3, 12 2011. Citado na página 18.