UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

IVAN ROBERTO STEIN

REDES NEURAIS LSTM NA MODELAGEM DE POLUENTES ATMOSFÉRICOS EM REGIÃO URBANA AGRÍCOLA

FRANCISCO BELTRÃO

2024

IVAN ROBERTO STEIN

REDES NEURAIS LSTM NA MODELAGEM DE POLUENTES ATMOSFÉRICOS EM REGIÃO URBANA AGRÍCOLA

LSTM neural networks in the modeling of atmospheric pollutants in an agricultural urban region

Dissertação de Mestrado apresentado como requisito para obtenção do título de Mestre em Engenharia Ambiental do Programa de pós-graduação em Engenharia Ambiental: Análise Tecnologia Ambiental da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Davi Zacarias de Souza

Coorientador: Prof. Dr. Claiton Zanini Brusamarello

FRANCISCO BELTRÃO

2024

4.0 Internacional

Esta licença permite download e compartilhamento do trabalho desde que sejam atribuídos créditos ao(s) autor(es), sem a possibilidade de alterá-lo ou utilizá-lo para fins comerciais. Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Francisco Beltrão

IVAN ROBERTO STEIN

REDES NEURAIS LSTM NA MODELAGEM DE POLUENTES ATMOSFÉRICOS EM REGIÃO URBANA AGRÍCOLA

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Engenharia Ambiental: Análise E Tecnologia Ambiental da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Meio Ambiente.

Data de aprovação: 05 de Setembro de 2024

Dr. Davi Zacarias De Souza, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Fernando Cavalcante Dos Santos, Doutorado - National University Of Singapore (Nus)

Dr. Fernando Cesar Manosso, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 05/09/2024.

AGRADECIMENTOS

A Deus, por me conceder força, sabedoria e a graça necessária para concluir mais esta etapa em minha vida.

Ao Professor Davi Zacarias de Souza, meu orientador, por sua orientação constante, paciência e apoio durante todo o processo, sem os quais este trabalho não seria possível.

À minha família, Liliane e Rafael, pela compreensão e apoio incondicional durante as inúmeras horas de ausência dedicadas a este projeto. Suas palavras de incentivo foram fundamentais.

Ao meu amigo Rodrigo de Moraes, que esteve ao meu lado desde o início, ajudando-me com motivação e nas nas preparações das apresentações. Sua amizade e apoio foram essenciais ao longo deste percurso.

Ao meu amigo Alcenir Vergílio Negri, cuja busca incessante por conhecimento foi uma verdadeira inspiração para que eu também me aprimorasse e alcançasse este objetivo.

Ao meu irmão Fernando Baggio Alcântara Oliveira, por sua generosidade em me auxiliar na correção deste trabalho, garantindo a qualidade do texto final.

Ao Professor Fernando Cavalcante dos Santos , por facilitar o acesso aos dados e fornecer orientações iniciais que foram cruciais para o desenvolvimento desta pesquisa.

A todos, o meu mais sincero e profundo agradecimento.

"A tecnologia é o conjunto de ferramentas que estende nossos horizontes, mas é a visão humana que transforma essas ferramentas em progresso."
Autor Desconhecido

RESUMO

A poluição do ar representa um desafio global com impactos severos na saúde pública e no meio ambiente, especialmente em áreas urbanas em crescimento como Francisco Beltrão, no Paraná, Brasil. Este estudo avaliou a eficácia das Redes Neurais Artificiais, especificamente o modelo *Long Short-Term Memory* (LSTM), na previsão das concentrações locais de poluentes atmosféricos como formaldeído (HCHO), dióxido de nitrogênio (NO_2) e dióxido de enxofre (SO_2), integrando dados de satélite recuperados do Portal da NASA, do satélite Sentinel-5P TROPOMI e variáveis meteorológicas locais. O modelo LSTM demonstrou a capacidade de capturar as complexas relações temporais e não-lineares presentes nos dados, resultando em previsões promissoras para as concentrações desses poluentes, particularmente para NO_2, com um erro quadrático médio (MSE) de $3{,}7885 \times 10^{-10}$. A análise revelou a influência significativa das condições meteorológicas na dinâmica da poluição do ar, destacando a importância de incorporá-las aos modelos preditivos. Os resultados também enfatizaram o impacto notável da temperatura na concentração de HCHO. Apesar das limitações, como a falta de dados de monitoramento no centro da cidade, este estudo fornece uma base embrionária para o desenvolvimento de ferramentas eficazes para monitorar e gerenciar a qualidade do ar em Francisco Beltrão, com aplicações potenciais em saúde pública, planejamento urbano e políticas ambientais. Esses resultados promissores abrem caminho para o desenvolvimento de sistemas de alerta e previsão da qualidade do ar, auxiliando na tomada de decisões e na implementação de políticas públicas eficazes para o gerenciamento da poluição do ar neste município.

Palavras-chave: poluição atmosferica; redes neurais lstm; qualidade do ar; previsão; francisco beltrão.

ABSTRACT

Atmospheric pollution represents a global challenge with severe impacts on public health and the environment, especially in growing urban areas like Francisco Beltrão in Paraná. This study evaluated the effectiveness of Artificial Neural Networks, specifically the Long Short-Term Memory (LSTM) model, in predicting local concentrations of atmospheric pollutants such as formaldehyde (HCHO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2), by integrating NASA satellite data with local meteorological variables. The LSTM model demonstrated the ability to capture the complex temporal and nonlinear relationships within the data, resulting in promising predictions for the concentrations of these pollutants, especially for NO concentration, with a mean squared error (MSE) de $3{,}7885 \times 10^{-10}$. Analysis of the results revealed the significant influence of meteorological conditions on the dynamics of atmospheric pollution, highlighting the importance of considering them in predictive models particularly the impact of temperature on the concentration of HCHO. Despite limitations such as the lack of monitoring data in the city center, this study provides an embryonic basis for the development of effective tools for monitoring and managing air quality in Francisco Beltrão, with potential applications in public health, urban planning, and environmental policies. These promising results pave the way for the development of air quality alert and forecasting systems, aiding decision-making and the implementation of effective public policies for managing atmospheric pollution in this municipality.

Keywords: atmospheric pollution; lstm neural networks; air quality; forecast; francisco beltrão.

LISTA DE FIGURAS

Figura 1 – Representação esquemática de um estrutura de Rede Neural Biológica 17

Figura 2 – Esquema representativo de uma rede com os dados de entrada, suas camadas e dados de saída 18

Figura 3 – Fluxo de Processamento e Componentes Internos de uma Célula *LSTM* 22

Figura 4 – Mapa Representativo da localização do Estado do Parana e da cidade de estudo, Francisco Beltrão 25

Figura 5 – Ponto Central de Francisco Beltrão-PR. 26

Figura 6 – Area Urbana de Francisco Beltrão 27

Figura 7 – Crescimento Populacional de Francisco Beltrão-PR. 28

Figura 8 – Frota de Veículos de Francisco Beltrão-PR. 28

Figura 9 – Serie temporal HCHO 39

Figura 10 – Serie temporal SO_2 40

Figura 11 – Serie temporal NO_2 42

Figura 12 – Comparação entre os valores reais e preditos de concentração de HCHO(mol/m^2) 43

Figura 13 – Comparação entre os valores reais e preditos de concentração de NO_2 (mol/m^2) 44

Figura 14 – Comparação entre os valores reais e preditos de concentração de SO_2(mol/m^2) 44

LISTA DE TABELAS

Tabela 1 – Equações: MAE, RMSE e MAPE 35

Tabela 2 – Dados de Concentração Reais de Gases mol/m^2 e Variáveis Meteorológicas 41

Tabela 3 – Dados de Predição de Concentração de Gases (mol/m^2) e Variáveis Meteorológicas 43

Tabela 4 – Métricas de Erro para Predições de Concentração de Gases 45

LISTAGEM DE CÓDIGOS FONTE

Listagem 1 – Importação de Bibliotecas **57**
Listagem 2 – Carregamento e Preparação dos Dados **57**
Listagem 3 – Normalização dos Dados **57**
Listagem 4 – Criação de Sequências **57**
Listagem 5 – Comprimento da Sequência **57**
Listagem 6 – Divisão dos Dados **57**
Listagem 7 – Criação de Sequências para Treinamento e Validação: **58**
Listagem 8 – Construção do Modelo LSTM **58**
Listagem 9 – Compilação do Modelo **58**
Listagem 10 – Treinamento do Modelo **58**
Listagem 11 – Avaliação do Modelo **58**
Listagem 12 – Previsões para o Período de Validação **58**
Listagem 13 – Desnormalização das Previsões **59**
Listagem 14 – Plotagem das Previsões Gerais **59**
Listagem 15 – Plotagem de Gráficos Individuais **59**

SUMÁRIO

1 INTRODUÇÃO 11

2 OBJETIVOS 12

2.1 Objetivo Geral 12

2.2 Objetivos Específicos 12

3 REVISÃO DA LITERATURA 13

3.1 Poluição Atmosférica 13

3.2 Impactos Ambientais dos Gases Poluentes 13

3.2.1 Formaldeído - HCHO 14

3.2.2 Dióxido de Nitrogênio - NO_2 14

3.2.3 Dióxido de Enxofre - SO_2 14

3.3 Importância das Variáveis Meteorológicas na Poluição Atmosférica . . . 15

3.3.1 Temperatura 15

3.3.2 Umidade Relativa 15

3.3.3 Velocidade do Vento 15

3.3.4 Precipitação 16

3.3.5 Radiação Solar 16

3.4 Redes Neurais 16

3.4.1 Neurônios Artificiais Unidades Fundamentais da Rede 17

3.4.2 Camadas: Organização Hierárquica da Rede 18

3.4.3 Função de Ativação: Introduzindo Não Linearidade 19

3.5 Redes Neurais Artificiais (RNAs) e Long Short-Term Memory (LSTM) . . 19

3.5.1 Redes Neurais Feedforward 20

3.5.2 Redes Neurais Recorrentes (RNNs) 20

3.5.3 Redes LSTM (Long Short-Term Memory) 20

3.5.4 Aplicações das Redes Neurais LSTM 22

3.6 Comparação com Métodos Tradicionais e Técnicas Modernas de Previsão de Poluentes Atmosféricos 23

3.7 Linguagem Python: Análise de Dados e Modelagem Ambiental 24

4 MATERIAIS E MÉTODOS 25

4.0.1 Definição do Problema e Requisitos 25

4.0.2 O Municipio: Francisco Beltrão 26
4.1 Dados Estatísticos 28
4.2 Ferramentas de Software 29
4.3 Modelagem Preditiva 30
4.3.1 Análise Exploratória dos Dados 32
4.3.2 Limpeza dos Dados 33
4.3.3 Divisão dos Dados 33
4.3.4 Criação de Sequências Temporais 33
4.3.5 Pré-processamento dos Dados 34
4.3.6 Desenvolvimento do Modelo LSTM 34
4.3.7 Treinamento e Otimização do Modelo 35
4.3.8 Avaliação do Desempenho do Modelo 35
5 RESULTADOS E DISCUSSÕES 36
5.1 Escolha e Justificativa da Arquitetura da Rede Neural 36
5.1.1 Manuseio de Dados Não Lineares 37
5.1.2 Treinamento e Validação 37
5.2 Pré-processamento e Tratamento de Dados 38
5.3 Avaliação do Modelo 41
5.3.1 Interpretação dos Resultados 45
5.4 Aplicações Práticas e Impactos 46
6 DISCUSSÕES 48
6.1 Limitações do Estudo 48
6.2 Desafios Futuros 48
7 CONCLUSÃO 50
7.1 Resultados e Contribuições 50
REFERÊNCIAS 52

ANEXOS 55
ANEXO A – CÓDIGOS FONTE 57

1 INTRODUÇÃO

A poluição do ar representa um desafio mundial que afeta seriamente tanto a saúde pública quanto o meio ambiente. Mesmo quando os níveis de poluentes são tidos como seguros, há uma ligação direta com diversas doenças graves, como derrame, problemas cardíacos, câncer de pulmão e dificuldades respiratórias. Esse quadro alarmante, reforçado pela intensificação dos episódios de poluição em várias cidades, levou a Organização Mundial da Saúde (OMS) a realizar, em 2016, uma análise global sobre os riscos da poluição do ar para a saúde. A crescente disponibilidade de dados sobre a qualidade do ar e a rápida evolução da ciência relacionada aos impactos da poluição na saúde reforçam a importância e a urgência de ações para mitigar esse problema global (WHO, 2000) .

Na América Latina, a poluição do ar tem sido apontada como um fator determinante no aumento da mortalidade por doenças pulmonares, cerebrovasculares e cardíacas, evidenciando a necessidade urgente de ações para melhorar a qualidade do ar (SANTOS, 2021). No Brasil, a Resolução CONAMA nº 491/2018 estabelece padrões de qualidade do ar, definindo os limites aceitáveis para diversos poluentes atmosféricos, emitidos por fontes fixas, como indústrias, e móveis, como veículos automotores (AMBIENTE, 2018).

Em áreas urbanas como Francisco Beltrão, no sudoeste do Paraná, a crescente população e a intensificação das atividades industriais e de transporte exercem pressão sobre a qualidade do ar. Com uma população estimada no último censo, em 2022, de 96.666 habitantes (IBGE, 2023), a cidade demanda soluções promissores para monitorar e gerenciar a poluição atmosférica, visando proteger a saúde pública e o meio ambiente.

Este estudo propõe a aplicação da rede neural *Long Short-Term Memory* (LSTM), um tipo de Rede Neural Artificial (RNA) de aprendizado profundo, para a previsão e o monitoramento da poluição atmosférica em Francisco Beltrão. As RNAs, modelos computacionais inspirados no sistema nervoso humano têm demonstrado grande potencial na previsão de fenômenos complexos e não lineares, como a poluição atmosférica. A arquitetura única do LSTM, com suas camadas de entrada, oculta, recorrente e de saída, permite a captura de dependências de longo prazo em séries temporais, tornando-o promissor na integração de dados de diversas fontes, como estações meteorológicas, satélites e modelos de dispersão de poluentes (Yuan *et al.*, 2018). A utilização do LSTM neste estudo visa apresentar mais uma alternativa de previsões e fornecer informações para o monitoramento da qualidade do ar em Francisco Beltrão.

As observações/medições de monitoramento dos gases poluentes estudados em Francisco Beltrão, considerando as características geográficas e meteorológicas específicas da região, promove uma técnica alternativa de avaliação dos impactos locais das emissões atmosféricas.

2 OBJETIVOS

2.1 Objetivo Geral

Aplicação das Redes Neurais Artificiais utilizando o modelo *Long Short-Term Memory (LSTM)* na predição de tendências das concentrações de gases poluentes atmosféricos na região urbana de Francisco Beltrão.

2.2 Objetivos Específicos

Coletar e processar dados do satélite **Sentinel-5 Precursor** ou abreviadamente **Sentinel-5P** sobre as concentrações atmosféricas de formaldeído (HCHO), dióxido de nitrogênio (NO_2) e dióxido de enxofre (SO_2) na região de Francisco Beltrão-PR.

Correlacionar os dados dos gases poluentes atmosféricos com parâmetros meteorológicos relevantes (temperatura, umidade, velocidade do vento).

Desenvolver modelo preditivo utilizando arquiteturas de Redes Neurais Artificiais, especificamente modelo *"Long Short-Term Memory"*, otimizando parâmetros para melhorar as previsões das concentrações de HCHO, NO_2 e SO_2 na região de estudo.

Validar o desempenho do modelo LSTM por meio de comparações com os dados observados, utilizando métricas de avaliação apropriadas, como o erro quadrático médio (MSE), para garantir a eficácia das previsões durante o período de validação.

Analisar os resultados das previsões e identificar padrões e tendências significativas entre as concentrações de poluentes na qualidade do ar em Francisco Beltrão-PR.

3 REVISÃO DA LITERATURA

3.1 Poluição Atmosférica

A poluição atmosférica, um problema global com graves consequências para a saúde humana e o meio ambiente é caracterizada pela presença de substâncias nocivas na atmosfera em concentrações que podem prejudicar a qualidade de vida, mesmo quando dentro dos limites estabelecidos por órgãos reguladores (AMÂNCIO; NASCIMENTO, 2012). Essas substâncias, originadas tanto de processos naturais quanto de atividades humanas, podem ser classificadas em poluentes primários, emitidos diretamente das fontes, e poluentes secundários, formados por reações químicas na atmosfera (CHAVES *et al.*, 2013; SILVA, 2015).

Eventos naturais como erupções vulcânicas e queimadas em biomassa, historicamente relevantes na contaminação atmosférica (CANÇADO *et al.*, 2006), ainda desempenham um papel na poluição do ar. No entanto, as atividades antrópicas, impulsionadas pelo rápido crescimento industrial e pela urbanização acelerada, tornaram-se as principais fontes dominantes de poluição, especialmente em áreas urbanas (TAETS, 2014). A queima de combustíveis fósseis para geração de energia, os processos industriais e o tráfego de veículos são exemplos de atividades que liberam na atmosfera uma variedade de poluentes, incluindo os gases formaldeído (HCHO), dióxido de nitrogênio (NO_2) e dióxido de enxofre (SO_2), foco deste estudo.

Os impactos da poluição atmosférica na saúde pública são alarmantes, variando de problemas respiratórios agudos a doenças crônicas como asma, bronquite, pneumonia, enfisema pulmonar e câncer de pulmão (PESTANA, 2015). A exposição prolongada a esses poluentes pode comprometer a qualidade de vida e aumentar a morbidade e a mortalidade, especialmente em grupos vulneráveis como crianças, idosos e pessoas com doenças preexistentes. Além dos efeitos na saúde, a poluição atmosférica contribui para a degradação ambiental, afetando a visibilidade, a vegetação, os ecossistemas aquáticos e o clima global (COSTA *et al.*, 2009).

Diante desse cenário, a implementação de políticas públicas promissoras para controlar e reduzir a poluição atmosférica é urgente e essencial. O monitoramento da qualidade do ar, o desenvolvimento de tecnologias limpas, a promoção de fontes de energia renováveis e a conscientização da população são algumas das medidas necessárias para proteger a saúde humana e o meio ambiente, e garantir um futuro mais sustentável para as próximas gerações.

3.2 Impactos Ambientais dos Gases Poluentes

Os gases poluentes como formaldeído (HCHO), dióxido de nitrogênio (NO_2) e dióxido de enxofre (SO_2) são responsáveis por diversos impactos ambientais, com efeitos que se estendem da saúde humana à degradação de ecossistemas. A presença desses poluentes na atmosfera,

provenientes de fontes antropogênicas e naturais, contribui para a deterioração da qualidade do ar, especialmente em áreas urbanas, e representa um desafio crescente para a sustentabilidade ambiental (BRITO; BERNARDO; GONÇALVES, 2022).

3.2.1 Formaldeído - HCHO

O formaldeído, um gás incolor com odor pungente, é liberado na atmosfera por diversas fontes, incluindo emissões industriais, combustão de combustíveis fósseis e até mesmo a decomposição de materiais orgânicos. A exposição ao HCHO pode causar irritação nos olhos, nariz e garganta, além de estar associada a problemas respiratórios, como asma e bronquite. Estudos também indicam que o formaldeído é um potencial carcinogênico humano, aumentando o risco de desenvolver câncer nas vias respiratórias superiores (AMÂNCIO; NASCIMENTO, 2012). A monitorização e o controle das emissões de HCHO são cruciais para proteger a saúde pública e minimizar seus impactos no meio ambiente.

3.2.2 Dióxido de Nitrogênio - NO_2

O dióxido de nitrogênio, um gás marrom-avermelhado com odor forte, é um dos poluentes atmosféricos em áreas urbanas. Originado especialmente da queima de combustíveis fósseis em veículos motorizados e processos industriais, o NO_2 contribui para a formação de ozônio troposférico, uma componente chave do *smog fotoquímico*. A exposição ao NO_2 pode causar irritação pulmonar, agravar doenças respiratórias preexistentes, como asma e bronquite, e aumentar a suscetibilidade a infecções respiratórias (CHAVES *et al.*, 2013; SILVA, 2015). Além dos efeitos na saúde humana, o NO_2 também contribui para a acidificação do solo e da água, prejudicando a vegetação e os ecossistemas aquáticos.

3.2.3 Dióxido de Enxofre - SO_2

O dióxido de enxofre, um gás incolor com odor pungente, é emitido principalmente pela queima de combustíveis fósseis que contêm enxofre, como carvão e óleo combustível, e por processos industriais como a produção de ácido sulfúrico. Na atmosfera, o SO_2 pode reagir com outras substâncias e formar partículas finas, que são prejudiciais à saúde humana. A exposição a este gás pode causar irritação nos olhos, nariz e garganta, além de problemas respiratórios, como bronquite e asma (CANÇADO *et al.*, 2006). O SO_2 também é um dos precursores da chuva ácida, que causa acidificação do solo e da água, prejudicando florestas, plantações e ecossistemas aquáticos.

3.3 Importância das Variáveis Meteorológicas na Poluição Atmosférica

Além das emissões de gases poluentes, as condições meteorológicas desempenham um papel fundamental na dinâmica da poluição atmosférica. Variáveis como temperatura, umidade relativa do ar e velocidade do vento influenciam diretamente a dispersão, transporte, transformação química e remoção dos poluentes na atmosfera (GUO *et al.*, 2020). Compreender essas interações é imperativo para desenvolver modelos preditivos mais precisos e implementar estratégias promissoras de mitigação da poluição do ar. A relação entre parâmetros meteorológicos e poluição atmosférica é complexa e multifacetada, com diversos fatores influenciando a concentração, dispersão e transformação química dos poluentes na atmosfera. Compreender essa relação é crucial para desenvolver modelos preditivos precisos e implementar estratégias promissoras de gestão da qualidade do ar:

3.3.1 Temperatura

A temperatura do ar afeta a taxa de reações químicas que levam à formação de poluentes secundários, como o ozônio troposférico. Além disso, o aumento da temperatura pode intensificar a evaporação de compostos orgânicos voláteis (COVs), que são precursores de ozônio e material particulado. A temperatura do ar desempenha um papel fundamental na dinâmica da poluição atmosférica. Temperaturas mais elevadas podem aumentar a volatilidade de alguns compostos, intensificando a emissão de poluentes como o formaldeído (HCHO). Além disso, o ar quente tende a subir, levando os poluentes para altitudes mais elevadas, onde podem ser transportados por longas distâncias. No entanto, em condições de inversão térmica, o ar frio fica aprisionado próximo à superfície, impedindo a dispersão dos poluentes e resultando em episódios de alta poluição (MACEDO, 2012).

3.3.2 Umidade Relativa

A umidade relativa do ar influencia a formação de partículas secundárias, como sulfatos e nitratos, que são componentes importantes do material particulado fino. Além disso, a umidade pode afetar a deposição úmida de poluentes, removendo-os da atmosfera através da chuva e da neve.

3.3.3 Velocidade do Vento

A velocidade e a direção do vento são fatores determinantes na dispersão e no transporte de poluentes. Ventos fortes podem diluir os poluentes e transportá-los para longe das fontes de emissão, enquanto ventos fracos podem levar à acumulação de poluentes em áreas

específicas. A velocidade e a direção do vento são fatores cruciais na dispersão dos poluentes atmosféricos. Ventos fortes favorecem a diluição e o transporte dos poluentes, reduzindo sua concentração em áreas próximas às fontes de emissão. Por outro lado, ventos fracos ou a ausência de vento podem levar à acumulação de poluentes, resultando em episódios de má qualidade do ar (SILVEIRA, 2022).

3.3.4 Precipitação

A chuva e a neve podem atuar como mecanismos de remoção de poluentes da atmosfera. As gotas de chuva capturam partículas e gases poluentes, depositando-os na superfície. Esse processo, conhecido como deposição úmida, contribui para a melhoria da qualidade do ar, especialmente após períodos de alta poluição (Vicentini, 2011). No entanto, a precipitação também pode levar à formação de ácidos na atmosfera, como o ácido sulfúrico e o ácido nítrico, que contribuem para a chuva ácida e seus impactos negativos no meio ambiente.

3.3.5 Radiação Solar

A radiação solar desempenha um papel importante na formação de poluentes secundários, como o ozônio troposférico. A luz solar desencadeia reações fotoquímicas entre os poluentes precursores, como óxidos de nitrogênio (NOx) e compostos orgânicos voláteis (COVs), levando à formação de ozônio e outros oxidantes (VALDAMBRINI; RIBEIRO, 2021).

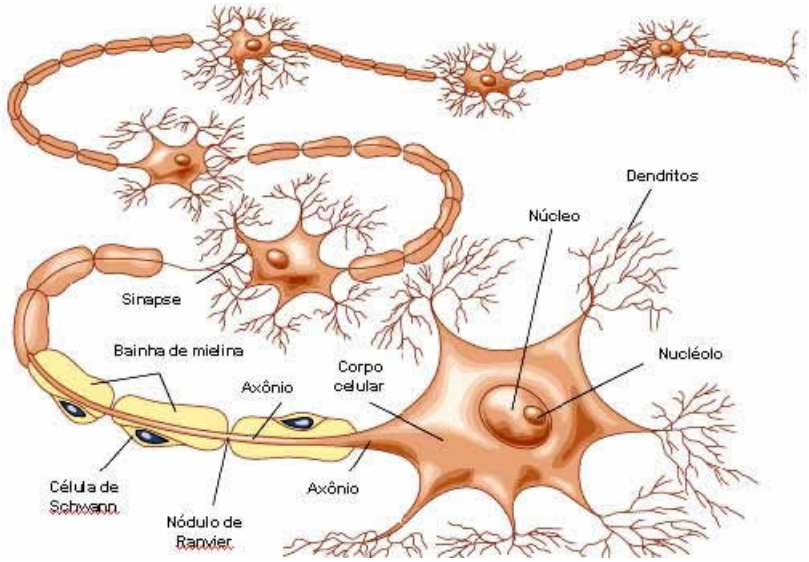
3.4 Redes Neurais

De acordo com Kaplan e Haenlein (2019), a inteligência artificial é descrita como a habilidade de um sistema em compreender de maneira precisa informações vindas de fontes externas, aprender com essas informações e aplicar esses conhecimentos para atingir metas e realizar tarefas específicas por meio de uma adaptação flexível.

Nesse contexto, as Redes Neurais Artificiais (RNA) imitam o funcionamento das redes neurais biológicas, que consistem em uma complexa rede de neurônios interconectados, como ilustrado na Figura 1. Cada neurônio recebe sinais através de suas dendrites, processa essas informações em seu núcleo e, em seguida, transmite uma resposta ao longo de seu axônio até as sinapses, que conectam o neurônio a outros neurônios.

Assim, compreender como se dá a criação e o uso dos neurônios artificiais, bem como a capacidade das suas aplicações em redes, é essencial para entender o desenvolvimento da IA. A Figura 1 demonstra uma representação esquemática dessa estrutura biológica, que serve como inspiração para o funcionamento das redes neurais artificiais.

Figura 1 – Representação esquemática de um estrutura de Rede Neural Biológica



Fonte: Adaptado de https://www.sobiologia.com.br/conteudos/FisiologiaAnimal/nervoso2.php (2024).

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados na estrutura e no funcionamento do sistema nervoso humano. Assim como o cérebro, as RNAs são capazes de aprender a partir de dados e realizar tarefas complexas, como reconhecimento de padrões, classificação e previsão(RAMOS, 2003).

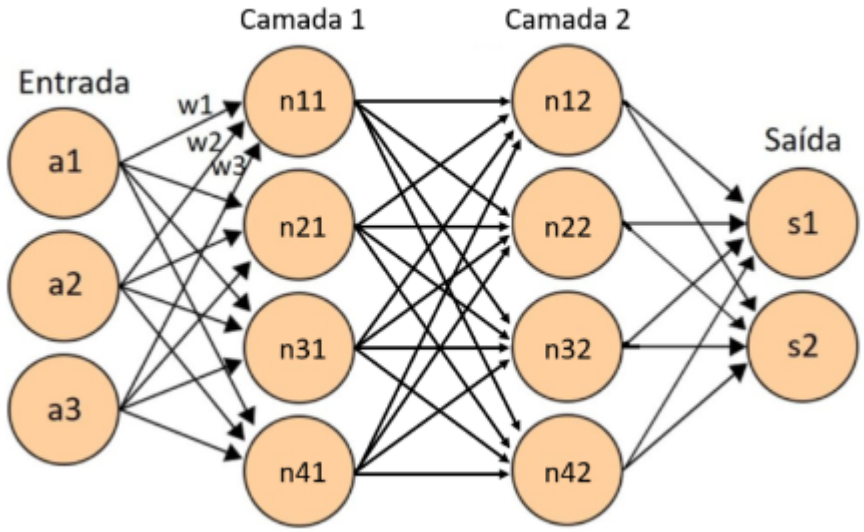
Na Figura 2, observamos a arquitetura básica de uma Rede Neural Artificial, que é composta por diferentes camadas: a camada de entrada, as camadas ocultas (Camada 1 e Camada 2) e a camada de saída. Cada círculo representa um neurônio artificial, que recebe sinais da camada anterior, processa esses sinais e os passa adiante, até gerar uma saída. As conexões entre os neurônios, representadas pelas linhas, possuem pesos (w1, w2, etc.) que são ajustados durante o processo de aprendizado, similarmente ao que ocorre nas sinapses das redes neurais biológicas (AVELINO *et al.*, 2022).

Portanto, as RNAs são sistemas de processamento de informação que imitam características de desempenho das redes neurais biológicas, permitindo a realização de tarefas computacionais avançadas, como mostrado na Figura 2.

3.4.1 Neurônios Artificiais Unidades Fundamentais da Rede

Os neurônios artificiais, também conhecidos como unidades de processamento, formam os componentes básicos das Redes Neurais Artificiais (RNAs). Cada neurônio artificial recebe um conjunto de sinais de entrada, cada um associado a um peso sináptico específico. Esses pesos sinápticos representam a intensidade da conexão entre os neurônios e são ajustados ao longo do processo de aprendizado da rede para refinar a resposta do modelo (SUN *et al.*, 2020).

Figura 2 – Esquema representativo de uma rede com os dados de entrada, suas camadas e dados de saída



Fonte: Adaptado de https://didatica.tech/introducao-a-redes-neurais-e-deep-learning/ (2024) .

Dentro de cada neurônio, as entradas são somadas de forma ponderada e, em seguida, é aplicada uma função de ativação. Esta função é crucial porque introduz não-linearidade ao processo, permitindo que a rede realize tarefas complexas e não apenas operações lineares simples. A função de ativação é o que determina a saída final do neurônio, influenciando diretamente como as informações são transmitidas ao longo da rede (JAGTAP; KAWAGUCHI; KARNIADAKIS, 2020).

3.4.2 Camadas: Organização Hierárquica da Rede

As RNAs são organizadas em camadas, que podem ser divididas em três tipos principais:

- Camada de Entrada: Recebe os dados de entrada e os distribui para os neurônios da primeira camada oculta.
- Camadas Ocultas: Realizam o processamento dos dados, extraindo características e padrões relevantes. O número de camadas ocultas e o número de neurônios em cada camada variam de acordo com a complexidade do problema e a arquitetura da rede.
- Camada de Saída: Produz a saída final da rede, que pode ser uma classificação, uma previsão ou um valor contínuo, dependendo da tarefa em questão.

3.4.3 Função de Ativação: Introduzindo Não Linearidade

A função de ativação é um componente crucial dos neurônios artificiais, responsável por determinar a saída do neurônio com base na soma ponderada das entradas. Ela introduz não linearidade ao modelo, permitindo que a rede neural aprenda relações complexas entre os dados de entrada e saída (RIBEIRO, 2020). As funções de ativação mais comuns incluem:

- Sigmóide: Gera uma saída suave e contínua entre 0 e 1, sendo particularmente útil em problemas de classificação binária. No entanto, é suscetível ao problema do *"vanishing gradient"* em redes profundas.
- Tanh (Tangente Hiperbólica): Similar à sigmóide, mas com a saída variando entre -1 e 1, o que pode ajudar a centralizar os dados em torno de zero. É frequentemente usada em problemas de regressão, mas também pode sofrer com o *"vanishing gradient"*.
- ReLU (Rectified Linear Unit): Uma função simples e eficaz que retorna 0 para entradas negativas e o próprio valor da entrada para entradas positivas. A ReLU é amplamente utilizada em redes neurais profundas devido à sua eficiência no treinamento, embora possa sofrer do problema de "neurônios mortos".
- Softmax: Utilizada principalmente na camada de saída em problemas de classificação multiclasse, ela transforma as saídas em uma distribuição de probabilidades sobre as classes possíveis, permitindo a interpretação direta das probabilidades.

A escolha da função de ativação depende da tarefa em questão e da arquitetura da rede. A não linearidade introduzida pela função de ativação permite que as RNAs aprendam relações complexas entre os dados de entrada e saída, tornando-as ferramentas poderosas para uma variedade de aplicações.

Modelos de redes neurais recorrentes têm sido aplicados na indústria e superam as redes neurais artificiais padrão em previsões, mas falham em previsões de séries temporais de longo prazo devido ao problema do gradiente que desaparece, apresentado como vanishing gradient (ABDULKADIR *et al.*, 2018).

Ainda segundo o mesmo autor, o modelo de Memória de Longo Curto Prazo (Long Short Term Memory - LSTM) visa superar o problema do gradiente que desaparece. O LSTM é especialmente projetado para evitar o problema de dependência de longo prazo, sendo este seu comportamento padrão.

3.5 Redes Neurais Artificiais (RNAs) e Long Short-Term Memory (LSTM)

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados na estrutura e no funcionamento do cérebro humano, capazes de aprender a partir de dados e realizar

tarefas complexas, como reconhecimento de padrões, classificação e previsão. Dentro do vasto universo das RNAs, existem dois tipos principais: as Redes Neurais Feedforward e as Redes Neurais Recorrentes (SARKER, 2021).

3.5.1 Redes Neurais Feedforward

Ainda segundo (SARKER, 2021), nesse tipo de RNA, as informações fluem em uma única direção, da camada de entrada para a camada de saída, sem ciclos de retroalimentação. Elas são promissoras para tarefas que não envolvem dependências temporais, como reconhecimento de imagens e classificação de texto.

3.5.2 Redes Neurais Recorrentes (RNNs)

As RNNs possuem conexões retroalimentadas, permitindo que informações de etapas anteriores sejam levadas em consideração no processamento atual. Essa característica as torna ideais para lidar com dados sequenciais, como séries temporais, onde a ordem dos dados é importante. No entanto, as RNNs tradicionais sofrem do problema do gradiente que desaparece ou explode, o que limita sua capacidade de aprender dependências de longo prazo (ABIODUN *et al.*, 2019).

3.5.3 Redes LSTM (Long Short-Term Memory)

As Redes LSTM são um tipo especial de RNN projetado para superar o problema do gradiente que desaparece ou explode. Segundo (SHERSTINSKY, 2020), as LSTM possuem uma estrutura interna mais complexa, com unidades de memória que permitem que informações relevantes sejam armazenadas e recuperadas ao longo do tempo:

- Mecanismo de Memória: A unidade de memória LSTM é composta por três portas principais:
- Porta de Entrada (Input Gate): Controla o fluxo de novas informações para a célula de memória.
- Porta de Esquecimento (Forget Gate): Determina quais informações devem ser descartadas da célula de memória.
- Porta de Saída (Output Gate): Controla o fluxo de informações da célula de memória para a saída da LSTM. Essas portas são reguladas por funções de ativação sigmóide e tanh, que permitem um controle preciso do fluxo de informações. A capacidade de armazenar e recuperar informações ao longo do tempo torna as LSTMs particularmente

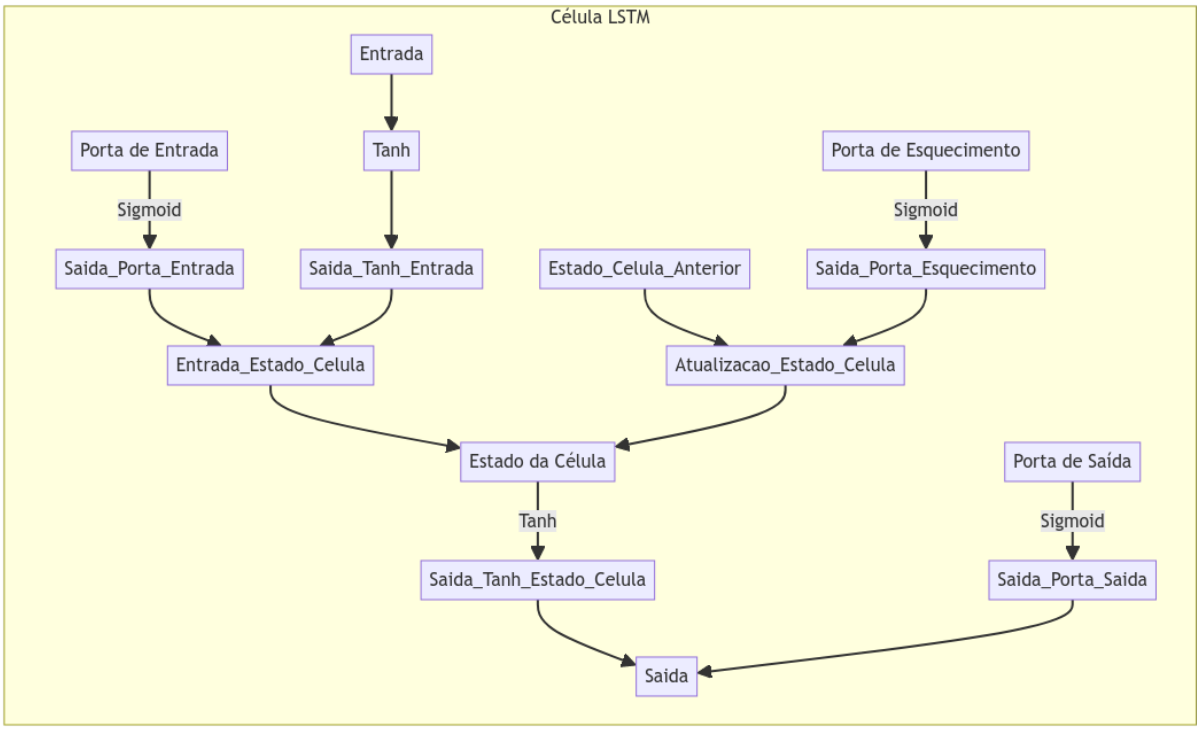
promissoras na modelagem de sequências longas e na captura de dependências temporais complexas(WANG *et al.*, 2020).

Essas portas são reguladas por funções de ativação sigmoide e tanh, que permitem um controle preciso do fluxo de informações. A capacidade de armazenar e recuperar informações ao longo do tempo torna as LSTMs particularmente promissoras na modelagem de sequências longas e na captura de dependências temporais complexas.

Na Figura 3, observam-se os aspectos:

- Célula LSTM: A unidade central da LSTM, responsável por armazenar o estado da célula.
- Portas: As três portas (entrada, esquecimento e saída) controlam o fluxo de informações para dentro e para fora da célula(SHERSTINSKY, 2020; ZHOU *et al.*, 2020).
- Fluxos de Informação: Setas indicam como as informações fluem entre a entrada, as portas, a célula e a saída da LSTM.
- Estados da Célula: Representa o estado interno da célula LSTM, que é atualizado a cada passo de tempo.
- Funções de Ativação: As portas utilizam funções sigmoide para controlar o fluxo de informações e a função tanh para processar as informações de entrada e o estado da célula.

Figura 3 – Fluxo de Processamento e Componentes Internos de uma Célula *LSTM*



Fonte: Autoria Própria (2024).

3.5.4 Aplicações das Redes Neurais LSTM

A LSTM têm sido amplamente utilizadas em diversas áreas (ABDULKADIR *et al.*, 2018), incluindo:

- Previsão de Séries Temporais: Previsão de demanda de energia, preços de ações, clima, etc.
- Processamento de Linguagem Natural: Tradução automática, reconhecimento de fala, análise de sentimentos, geração de texto.
- Reconhecimento de Voz: Transcrição de áudio, identificação de palavras-chave, sistemas de diálogo.
- Visão Computacional: Reconhecimento de objetos, rastreamento de movimento, análise de vídeo.

3.6 Comparação com Métodos Tradicionais e Técnicas Modernas de Previsão de Poluentes Atmosféricos

A previsão de poluentes atmosféricos pode ser realizada utilizando diferentes abordagens, que vão desde modelos estatísticos tradicionais até métodos mais recentes, baseados em aprendizado de máquina. Métodos tradicionais, como regressões lineares, médias móveis e autoregressive integrated moving average (ARIMA), são amplamente aplicados devido à sua simplicidade e capacidade de fornecer insights rápidos sobre tendências históricas dos dados (KUMAR; JAIN; KUMAR, 2016). No entanto, esses métodos apresentam limitações significativas, especialmente ao lidar com dados não lineares e complexos, como os envolvidos na previsão de poluentes atmosféricos, que são altamente influenciados por variáveis meteorológicas e emissões antropogênicas.

Por outro lado, técnicas modernas, como as redes neurais artificiais, especificamente a arquitetura Long Short-Term Memory (LSTM), têm demonstrado maior precisão na modelagem de dados temporais e não lineares, particularmente em séries temporais complexas (HOCHREITER; SCHMIDHUBER, 1997). Estudos recentes indicam que as LSTMs superam métodos tradicionais em termos de precisão e robustez quando aplicadas à previsão de poluentes atmosféricos, graças à sua capacidade de capturar dependências de longo prazo nas séries temporais (WANG *et al.*, 2020; ZHAO; ZHANG; XU, 2019).

Um estudo de Zhao et al. (2019), publicado no Atmospheric Environment, comparou a eficácia de métodos tradicionais e LSTM na previsão da qualidade do ar em grandes cidades da China. Os resultados mostraram que o LSTM obteve um desempenho superior, com uma redução significativa no erro quadrático médio (MSE) em comparação aos métodos ARIMA e médias móveis. Outro estudo, conduzido por (WANG; DI; ZHANG, 2020) e publicado no Journal of Cleaner Production, destacou que os modelos baseados em aprendizado profundo, como as redes neurais, conseguem capturar de maneira mais eficiente a complexa interação entre variáveis meteorológicas e concentrações de poluentes, fornecendo previsões mais precisas e com menos variabilidade em períodos de validação.

Além disso, as técnicas modernas possuem a vantagem de serem adaptáveis e de poderem ser continuamente treinadas com novos dados, o que aumenta a sua capacidade preditiva em cenários de mudanças rápidas, como as condições atmosféricas e eventos de emissão extrema (TANG; WANG, 2021). No entanto, esses modelos também exigem maior capacidade computacional e conhecimento técnico para implementação, em contraste com os métodos tradicionais, que, embora menos precisos, são mais fáceis de implementar e interpretar.

Dessa forma, enquanto os modelos tradicionais ainda têm valor em aplicações simples e de curto prazo, as técnicas modernas, especialmente as redes LSTM, estão emergindo como a escolha ideal para cenários que envolvem a previsão de poluentes atmosféricos em ambientes urbanos complexos (LIU *et al.*, 2019).

3.7 Linguagem Python: Análise de Dados e Modelagem Ambiental

Python, uma linguagem de programação de alto nível, interpretada e multiplataforma, destaca-se por sua sintaxe clara e legível, tornando-a acessível tanto para iniciantes quanto para programadores experientes. Sua tipagem dinâmica e a vasta biblioteca de módulos e frameworks a tornam uma ferramenta versátil e poderosa para diversas aplicações, incluindo a análise de dados e a modelagem ambiental.

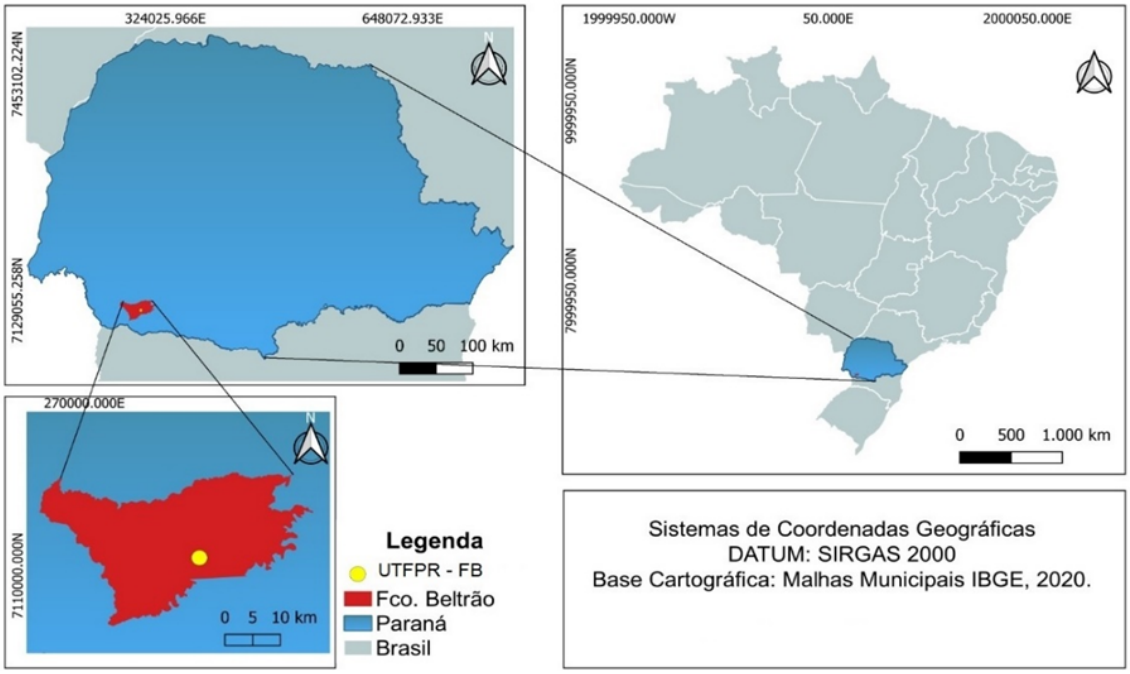
No contexto da pesquisa em questão, a linguagem Python oferece um ambiente propício para a manipulação e análise de dados geoespaciais e meteorológicos, bem como para o desenvolvimento e implementação de modelos de aprendizado de máquina, como as Redes Neurais Artificiais (RNAs) e as Redes LSTM, utilizadas neste estudo.

4 MATERIAIS E MÉTODOS

4.0.1 Definição do Problema e Requisitos

A primeira etapa envolveu a identificação do problema e a definição dos requisitos do sistema. A preocupação com a qualidade do ar em Francisco Beltrão,Figura 4, intensificada pelo aumento da população, da frota de veículos e das atividades industriais e agrícolas, demanda soluções promissoras para monitorar e prever a poluição atmosférica. Diante desse cenário, este estudo visa abordar o problema da poluição atmosférica na cidade, buscando desenvolver um sistema capaz de prever as concentrações de poluentes como Formaldeído (HCHO), Dióxido de Nitrogênio (NO_2) e Dióxido de Enxofre (SO_2).

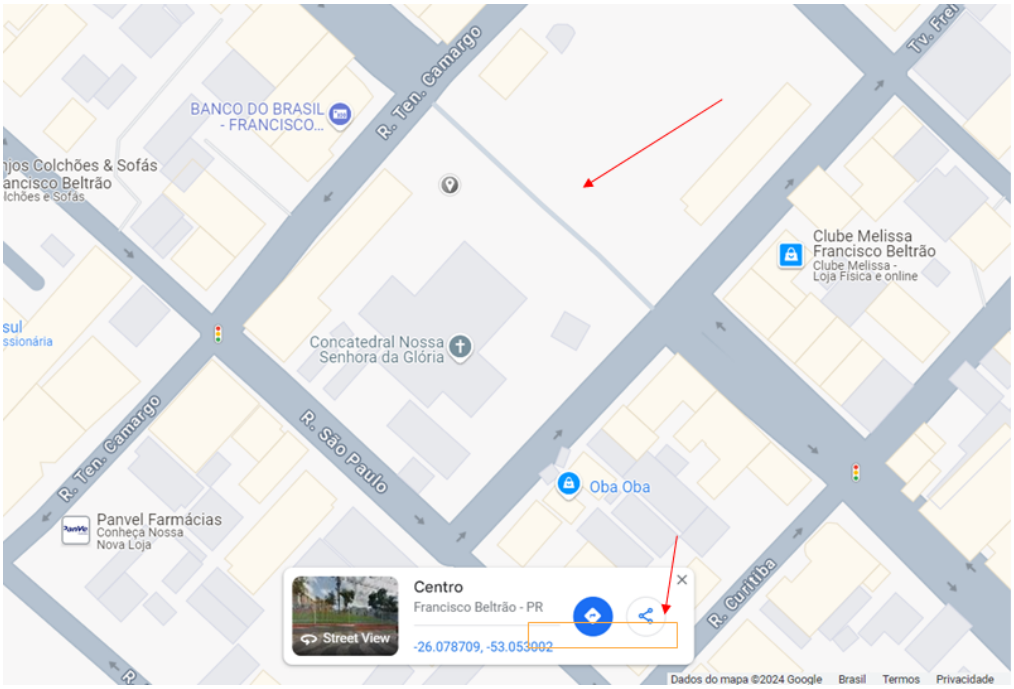
Figura 4 – Mapa Representativo da localização do Estado do Parana e da cidade de estudo, Francisco Beltrão



Fonte: Adaptado de (IBGE, 2023).

Para o alcance espacial, foram aplicados os seguintes parâmetros: latitude -26.078 e longitude -53.053. Este ponto corresponde a praça central do municipio conforme apresentado na Figura 5.

Figura 5 – Ponto Central de Francisco Beltrão-PR.



Fonte: Autoria Própria (2024).

4.0.2 O Municipio: Francisco Beltrão

Francisco Beltrão, Figura 6, é uma cidade localizada no sudoeste do estado do Paraná, Brasil. Situada a uma altitude média de 550 metros acima do nível do mar, a cidade possui um clima subtropical úmido, caracterizado por verões quentes e invernos amenos, com chuvas bem distribuídas ao longo do ano. Francisco Beltrão é cortada pelo Rio Marrecas, que desempenha um papel importante na geografia e economia local , contribuindo para a agricultura e o abastecimento de água. A vegetação predominante na região é a Mata Atlântica, embora grande parte da cobertura original tenha sido substituída por áreas agrícolas e pastagens (FARIAS; MENDONCA, 2022).

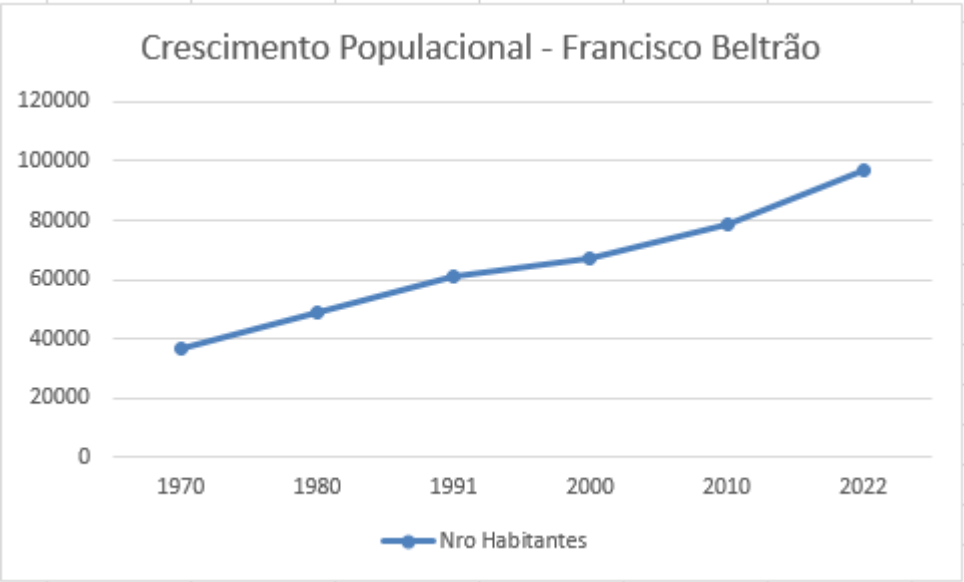
Figura 6 – Area Urbana de Francisco Beltrão

Fonte: Adaptado de https://especiais.gazetadopovo.com.br/ (2024).

Economicamente, Francisco Beltrão é um importante centro regional, com destaque para a agropecuária, especialmente a produção de soja, milho, e a criação de suínos e aves. O setor de serviços também é significativo, impulsionado pelo comércio e pela prestação de serviços de saúde e educação. A cidade conta com uma infraestrutura desenvolvida, incluindo universidades, hospitais, e uma rede de transporte que facilita o acesso às cidades vizinhas e aos mercados consumidores (COSTA; SOUZA, 2020).

O contínuo crescimento populacional de Francisco Beltrão (Figura 7) e a melhoria de renda tem apresentado um aumento significativo da frota de veículos e, consequentemente, da poluição do ar. A concentração urbana intensifica as emissões de poluentes atmosféricos, desafiando o controle da qualidade do ar e a preservação da qualidade de vida.

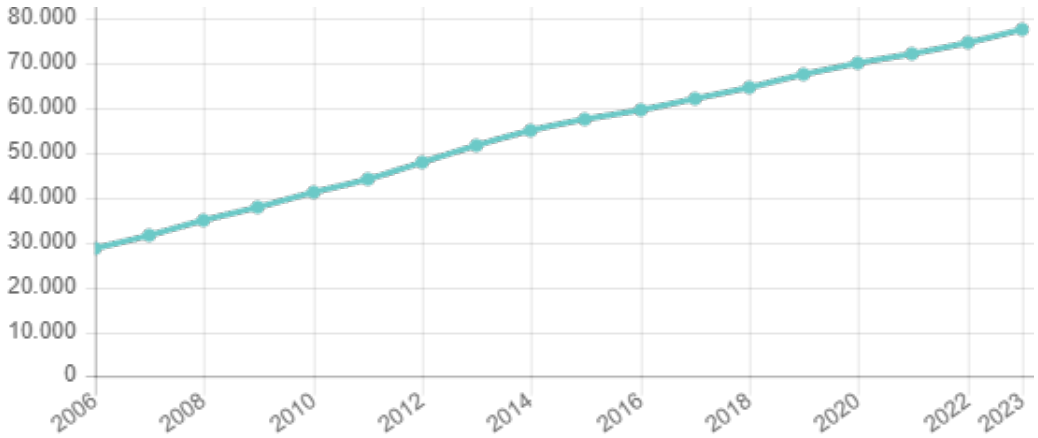
Figura 7 – Crescimento Populacional de Francisco Beltrão-PR.



Fonte: Adaptado de IBGE (2023).

Francisco Beltrão, com 96.666 habitantes(IBGE, 2023) e uma frota de 74.617 veículos, Figura 8, enfrenta desafios adicionais devido à presença de atividades agrícolas e industriais, que contribuem para a emissão de poluentes(IBGE - Instituto Brasileiro de Geografia e Estatística, 2022). A poluição do ar afeta a saúde da população, especialmente daqueles com problemas respiratórios, e causa impactos negativos no meio ambiente.

Figura 8 – Frota de Veículos de Francisco Beltrão-PR.



Fonte: Adaptado de IBGE - Instituto Brasileiro de Geografia e Estatística (2022).

4.1 Dados Estatísticos

Os dados são a base fundamental desta pesquisa, provenientes de fontes específicas e com características temporais. A coleta de dados foi realizada a partir de duas principais fontes:

Portal NASA: Durante a coleta de dados no Portal da NASA (https://disc.gsfc.nasa.gov/), foi utilizado um conjunto específico de filtros para obter infor-

mações relevantes para o estudo. O período selecionado foi de 01/05/2018 a 31/10/2023, abrangendo 2009 dias de observação. Para o alcance espacial, foram aplicados os seguintes parâmetros: latitude -26.078 e longitude -53.053. Dentro dessas coordenadas, foram filtrados os dados de *Atmospheric Chemistry (650)*, com resolução espacial de 5,5 km x 3,5 km. O satélite selecionado para a coleta foi o *Sentinel-5P TROPOMI*, conhecido por sua alta resolução e capacidade de monitoramento de gases atmosféricos.

Os downloads realizados resultaram em um arquivo por dia para cada um dos três gases diferentes monitorados, totalizando 6.027 arquivos. Cada arquivo tinha um tamanho aproximado de 697 MB, acumulando um total de 4.200.819 MB, ou aproximadamente 4,2 *terabytes* de informação. Este grande volume de dados proporciona uma base robusta para análises detalhadas sobre a composição atmosférica e as variações espaciais e temporais dos gases na região especificada, oferecendo dados valiosos para estudos ambientais e climáticos.

As informações sobre as variáveis meteorológicas foram obtidos da Estação Meteorológica da UTFPR - campus Francisco Beltrão.

4.2 Ferramentas de Software

As ferramentas de software foram selecionadas com o objetivo de processar, analisar e modelar os dados coletados. Abaixo estão as principais ferramentas utilizadas:

O software Visual Studio Code, um ambiente de desenvolvimento integrado (IDE) popular devido à sua flexibilidade, extensibilidade e suporte para diversas linguagens de programação. A escolha do Visual Studio Code foi motivada pela sua ampla documentação disponível na internet, o que facilita a resolução de problemas e a aprendizagem de novas funcionalidades.

Linguagem Python: Python foi escolhida por ser uma linguagem de programação poderosa e versátil, com uma vasta quantidade de bibliotecas disponíveis para análise de dados e aprendizado de máquina. As bibliotecas específicas utilizadas foram:

- **Pandas:** Utilizada para manipulação de dados, pré-processamento e estruturação dos conjuntos de dados temporais.
- **NumPy:** Essencial para operações matemáticas e manipulação de arrays, oferecendo suporte para operações numéricas de alta performance.
- **TensorFlow e Keras:** Frameworks essenciais para construção e treinamento de redes neurais artificiais. A biblioteca TensorFlow foi utilizada para implementar modelos de redes neurais, enquanto Keras facilitou a configuração e compilação desses modelos.
- **Scikit-learn:** Utilizada para pré-processamento de dados, normalização e separação de conjuntos de treinamento e validação. Esta biblioteca fornece ferramentas eficientes para análise de dados e aprendizado de máquina.

- **Matplotlib:** Utilizada para visualização de dados, permitindo a criação de gráficos e plots para análise exploratória dos dados.

A combinação dessas ferramentas permitiu um fluxo de trabalho eficiente, desde a coleta e pré-processamento dos dados até a modelagem e visualização dos resultados, garantindo uma análise abrangente e precisa dos dados coletados.

4.3 Modelagem Preditiva

A modelagem preditiva foi realizada com base em redes neurais LSTM (*Long Short-Term Memory*), uma arquitetura especializada para lidar com dados sequenciais. A estrutura do modelo incluiu:

Camadas LSTM: Duas camadas LSTM foram implementadas para capturar dependências temporais nos dados.

Camada Dense: Camada de saída para prever múltiplas variáveis climáticas, dimensionada de acordo com o número de características no conjunto de dados.

Para implementar essa estrutura, foram utilizadas diversas bibliotecas e ferramentas de software que facilitam o processamento e análise dos dados.

A seguir, é descrita cada parte do código utilizado para a modelagem preditiva:

Listagem de código 1 - Importação de Bibliotecas

A primeira etapa do processo consiste na importação das bibliotecas essenciais para a construção do modelo e a manipulação dos dados. A biblioteca *sys* é utilizada para configurar a codificação de saída, enquanto pandas e *numpy* são fundamentais para a manipulação e estruturação dos dados. O *tensorflow* é empregado para a criação do modelo *LSTM*, que será responsável pelas previsões. O *MinMaxScaler*, da biblioteca *sklearn*, é utilizado para normalizar os dados, garantindo que eles fiquem em uma escala apropriada para o treinamento da rede neural. Além disso, a biblioteca *matplotlib* é utilizada para a visualização gráfica dos resultados, o que facilita a análise do desempenho do modelo.

Listagem de código 2 - Carregamento e Preparação dos Dados:

Nesta etapa, os dados são carregados a partir de um arquivo *CSV*, e a coluna de datas é convertida para o formato datetime, sendo definida como o índice do *DataFrame*. Essa preparação é crucial para garantir que as séries temporais sejam manipuladas corretamente, permitindo a criação de sequências de dados para o treinamento da *LSTM*.

Listagem de código 3 - Normalização dos Dados

Os dados são normalizados utilizando o método *MinMaxScaler*, que transforma os valores em uma escala entre 0 e 1. Essa técnica de normalização é fundamental para evitar que variáveis com magnitudes muito diferentes interfiram no desempenho do modelo, melhorando a convergência durante o treinamento.

Listagem de código 4 - Criação de Sequências:

Para treinar o modelo *LSTM*, é necessário criar sequências de dados. Isso é feito por meio de uma função que divide os dados em janelas temporais de comprimento especificado, neste caso, 10 dias. Essas sequências permitem que o modelo capture as dependências temporais presentes nos dados, o que é essencial para a eficácia da LSTM em tarefas de previsão.

Listagem de código 5 - Definição do Comprimento da Sequência:

O comprimento das sequências de entrada é definido como 10 dias. Esse valor foi selecionado com base na análise das características do conjunto de dados e na necessidade do modelo de capturar padrões temporais de curto prazo.

Listagem de código 6 - Divisão dos Dados:

Os dados são divididos em dois subconjuntos: o conjunto de treinamento, que inclui os dados de 2018 a 2022, e o conjunto de validação, que compreende o período de 2023. Essa divisão é essencial para avaliar o desempenho do modelo em dados que não foram utilizados durante o treinamento, simulando um cenário de previsão real.

Listagem de código 7 - Criação de Sequências para Treinamento e Validação:

Tanto para o conjunto de treinamento quanto para o de validação, são criadas sequências de entrada e saída, com base no comprimento da sequência definido anteriormente. Essa etapa assegura que o modelo seja treinado e validado com dados consistentes e corretamente estruturados.

Listagem de código 8 - Construção do Modelo LSTM:

O modelo preditivo é construído utilizando uma arquitetura sequencial com duas camadas *LSTM*. A primeira camada contém 64 unidades e retorna sequências, enquanto a segunda camada, com 32 unidades, não retorna sequências. A última camada do modelo é uma camada densa (*Dense*) que gera a previsão final de acordo com o número de características dos dados. A estrutura da rede *LSTM* foi projetada para capturar relações temporais complexas nos dados de entrada..

Listagem de código 9 - Compilação do Modelo:

O modelo é compilado utilizando o otimizador Adam, amplamente utilizado em redes neurais devido à sua eficiência na atualização de pesos durante o treinamento. A função de perda utilizada é o erro quadrático médio (MSE), que é adequada para medir a diferença entre os valores reais e previstos em tarefas de regressão.

Listagem de código 10 - Treinamento do Modelo:

O treinamento do modelo ocorre em 50 épocas, com um tamanho de batch de 32. Durante o treinamento, os dados de validação são usados para monitorar o desempenho do modelo, permitindo ajustes e prevenindo o sobreajuste. Essa etapa é essencial para otimizar os parâmetros do modelo e assegurar que ele generalize bem para novos dados.

Listagem de código 11 - Avaliação do Modelo:

Após o treinamento, o modelo é avaliado nos conjuntos de dados de treinamento e validação, e as respectivas perdas (erros) são impressas. Essa avaliação oferece uma visão

quantitativa do desempenho do modelo, indicando o quão bem ele consegue prever os valores no conjunto de validação em comparação com o conjunto de treinamento.

Listagem de código 12 - Previsões para o Período de Validação:

Nesta etapa, são feitas previsões para o período de validação (21 a 31 de outubro de 2023). O modelo treinado é utilizado para prever os valores de saída com base nas sequências de entrada criadas anteriormente, fornecendo uma avaliação prática da sua capacidade preditiva..

Listagem de código 13 - Desnormalização das Previsões:

As previsões geradas pelo modelo são desnormalizadas, retornando à escala original dos dados. Esse processo é fundamental para interpretar corretamente os resultados, já que os valores normalizados não representam diretamente os dados reais.

Listagem de código 14 - Plotagem das Previsões Gerais:

Para facilitar a visualização dos resultados, é gerado um gráfico que compara os valores reais com os valores previstos pelo modelo para todas as variáveis climáticas no período de validação. A visualização gráfica é uma ferramenta poderosa para identificar padrões e possíveis discrepâncias entre os valores reais e previstos.

Listagem de código 15 - Plotagem de Gráficos Individuais: Adicionalmente, são gerados gráficos individuais para cada variável climática, comparando os valores reais e previstos. Essa abordagem detalhada permite uma análise mais granular do desempenho do modelo em cada variável específica.

Esta seção descreve como as ferramentas de software e bibliotecas foram utilizadas para a construção, treinamento e avaliação do modelo LSTM. Cada etapa do processo de modelagem preditiva é explicada em detalhes, enquanto os códigos correspondentes estão dispostos na seção de anexos, proporcionando uma visão organizada e clara da implementação.

4.3.1 Análise Exploratória dos Dados

A análise exploratória dos dados é uma etapa crucial em qualquer projeto de aprendizado de máquina, e neste estudo, não foi diferente. Ao receber os dados da NASA e da estação meteorológica da UTFPR, foi essencial realizar uma análise minuciosa para entender as características, padrões e tendências presentes nas series temporais. Essa análise inicial permitiu identificar possíveis problemas, como valores ausentes, *outliers* e inconsistências, que poderiam comprometer a qualidade do modelo preditivo.

Diversas técnicas foram empregadas nessa fase, incluindo a análise estatística descritiva, a visualização gráfica e a identificação de correlações entre as variáveis. A análise estatística descritiva forneceu informações sobre a distribuição dos dados, como média, mediana, desvio padrão e valores mínimo e máximo. Já a visualização gráfica, por meio de histogramas, gráficos de dispersão e gráficos de linha, permitiu identificar visualmente padrões e tendências nos dados.

A identificação de correlações entre as variáveis foi outro ponto importante da análise exploratória. Ao calcular a correlação entre os diferentes poluentes atmosféricos (HCHO, NO_2 e SO_2) e as variáveis meteorológicas (temperatura, umidade e velocidade do vento), foi possível identificar quais variáveis meteorológicas exercem maior influência na concentração dos poluentes. Essa informação foi fundamental para a seleção das variáveis de entrada do modelo LSTM, garantindo que o modelo utilize as informações mais relevantes para realizar as previsões.

4.3.2 Limpeza dos Dados

Durante o pré-processamento dos dados, identificou-se a presença de valores 'NaN' (Not a Number) nas séries temporais. Esses valores podem surgir devido a erros na geração dos dados ou falhas durante a coleta. A presença de 'NaN' pode comprometer a integridade da análise, introduzindo vieses ou distorções nos resultados. Portanto, optou-se por remover essas entradas das séries temporais para assegurar a qualidade dos dados utilizados no treinamento do modelo.

A literatura destaca a importância de lidar adequadamente com valores ausentes ou inválidos em séries temporais. (KAMALOV; SULIEMAN, 2021)realizaram um estudo comparativo sobre métodos de recuperação de sinais em séries temporais, enfatizando os impactos de diferentes técnicas de imputação na precisão dos dados, portanto, foi necessário remover estes valores das series.

4.3.3 Divisão dos Dados

Os dados foram divididos em conjuntos de treinamento e validação de forma estratificada para garantir que ambos representassem adequadamente o comportamento temporal dos dados. O conjunto de treinamento abrange o período de 1º de maio de 2018 a 31 de dezembro de 2022, enquanto o conjunto de validação cobre de 1º de janeiro de 2023 a 20 de outubro de 2023. Essa separação é crucial para avaliar a capacidade de generalização do modelo e evitar *overfitting*, assegurando que o modelo seja avaliado em dados que não foram utilizados durante o treinamento.

4.3.4 Criação de Sequências Temporais

Para treinar a rede LSTM, os dados foram transformados em sequências temporais. Este processo envolveu a criação de janelas de tempo que capturassem as dependências temporais entre os diferentes registros. Cada janela de tempo consistia em uma sequência de registros, que eram usadas como entradas para a LSTM. O registro subsequente à sequência servia como

o valor alvo a ser previsto. Dessa forma, a rede LSTM foi capaz de aprender e prever padrões temporais, aproveitando a estrutura sequencial dos dados.

4.3.5 Pré-processamento dos Dados

Após a análise exploratória, os dados passaram por um processo de pré-processamento, que visa preparar os dados para o treinamento do modelo LSTM. Nessa etapa, foram realizadas as seguintes atividades:

Tratamento de Valores Ausentes: Identificação e tratamento de valores ausentes nos dados, seja por meio da imputação com valores médios, medianos ou outros métodos adequados, ou pela remoção de linhas ou colunas com dados faltantes.

Tratamento de Outliers: Identificação e tratamento de outliers, que são valores extremos que podem distorcer o aprendizado do modelo. Os outliers podem ser removidos ou substituídos por valores mais adequados. Normalização dos Dados: Os dados foram normalizados para garantir que todas as variáveis tenham a mesma escala, evitando que algumas variáveis dominem o aprendizado do modelo devido a suas magnitudes. A normalização Min-Max foi utilizada para escalar os valores entre 0 e 1.

Criação de Sequências Temporais: Os dados foram organizados em sequências temporais, onde cada sequência contém um número fixo de observações passadas. Essas sequências são utilizadas como entrada do modelo LSTM, que utiliza as informações passadas para prever os valores futuros.

4.3.6 Desenvolvimento do Modelo LSTM

Com os dados devidamente pré-processados, a próxima etapa foi o desenvolvimento do modelo LSTM. A arquitetura do modelo foi definida com base na análise exploratória dos dados e na revisão da literatura, buscando otimizar a capacidade do modelo de capturar as dependências temporais e as relações complexas entre as variáveis.

O modelo LSTM foi implementado utilizando as bibliotecas TensorFlow e Keras, que fornecem ferramentas poderosas para a construção e o treinamento de redes neurais. A estrutura do modelo incluiu as seguintes camadas:

Camadas LSTM: Duas camadas LSTM foram utilizadas para capturar as dependências temporais nos dados. A primeira camada LSTM foi configurada para retornar sequências, permitindo que a informação temporal seja passada para a próxima camada. A segunda camada LSTM não retorna sequências, produzindo um único vetor de saída que representa a previsão para o próximo instante de tempo.

Camada Densa: Uma camada densa foi adicionada após as camadas LSTM para realizar a previsão final das concentrações de poluentes. Essa camada utiliza a função de ativação linear para produzir as previsões.

4.3.7 Treinamento e Otimização do Modelo

O treinamento do modelo LSTM foi realizado utilizando os dados de treinamento pré-processados. Durante o treinamento, o modelo ajusta seus parâmetros internos (pesos e vieses) para minimizar o erro entre suas previsões e os valores reais. O processo de treinamento envolveu a utilização de um otimizador (Adam) e de uma função de perda (Erro Quadrático Médio), que medem a qualidade das previsões do modelo.

Para otimizar o desempenho do modelo, foram utilizados os dados de validação. A cada época de treinamento, o modelo era avaliado nos dados de validação, e o erro de validação era monitorado. O objetivo era encontrar o ponto ideal de treinamento, onde o erro de validação começa a aumentar, indicando que o modelo está começando a superajustar aos dados de treinamento.

4.3.8 Avaliação do Desempenho do Modelo

Após o treinamento, o modelo LSTM foi avaliado nos dados de teste, que são dados que o modelo nunca viu antes. Essa etapa é crucial para verificar a capacidade do modelo de generalizar para novos dados e garantir que ele não esteja superajustado aos dados de treinamento. Para avaliar o desempenho do modelo, foram utilizadas 3 métricas, são elas: Erro Quadrático Médio (MSE), o Erro Absoluto Médio (MAE) e o Coeficiente de Determinação (R^2). Essas métricas fornecem informações sobre a precisão das previsões do modelo em relação aos valores reais. Na tabela 1 demonstra-se com são calculas estas métricas.

Tabela 1 – Equações: MAE, RMSE e MAPE

Sigla	Equação
MAE	$\frac{1}{n}\sum_{i=1}^{n}\|(Y_i - \hat{Y}_i)\|$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$
MAPE	$\frac{1}{n}\sum_{i=1}^{n}\frac{\|(Y_i - \hat{Y}_i)\|}{Y_i} \times 100$

Fonte: Adaptado de: (ADAMS *et al.*, 2015).

5 RESULTADOS E DISCUSSÕES

5.1 Escolha e Justificativa da Arquitetura da Rede Neural

Para realizar a previsão dos dados de concentração de dados foi escolhida uma arquitetura de rede neural recorrente, especificamente uma *Long Short-Term Memory* (LSTM). A escolha dessa arquitetura foi fundamentada em várias razões. A arquitetura da rede neural LSTM utilizada consistiu nas seguintes camadas:

Camada LSTM Inicial: Unidades: 64 - Retornar Sequências: Sim - Função de Ativação: tanh

A primeira camada LSTM, com um número relativamente alto de unidades (64), permite capturar relações temporais complexas nos dados. A configuração $return_sequences = True$ garante que a sequência completa de saídas para cada passo temporal seja passada para a próxima camada LSTM. Isso é fundamental para permitir que camadas subsequentes continuem a processar e refinar as informações temporais capturadas.

Camada LSTM Intermediária: Unidades: 32 - Retornar Sequências: Não - Função de Ativação: tanh

A segunda camada LSTM, com 32 unidades, reduz a dimensionalidade da sequência de saída, permitindo a extração de padrões temporais mais abstratos e refinados. A configuração $return_sequences = False$ transforma a sequência em uma única representação, agregando toda a informação temporal em uma única saída. Isso prepara os dados de maneira adequada para a camada densa final, que fará as previsões baseadas nesta representação condensada.

Camada Densa - Unidades: [Número de variáveis de saída] - Função de Ativação: Linear

A camada densa final é responsável por gerar as previsões para cada variável de saída. A função de ativação linear é apropriada para problemas de regressão, onde o objetivo é prever valores contínuos. Isso permite que a rede produza previsões precisas para cada variável de concentração de gases.

As redes LSTM são robustas contra problemas de gradiente de desaparecimento e explosão, que são comuns em redes neurais recorrentes tradicionais. Isso permite que a LSTM mantenha informações relevantes ao longo de longas sequências, melhorando a precisão das previsões (SHERSTINSKY, 2020).

A natureza multivariada das séries temporais de concentração de gases, onde diversos sensores capturam simultaneamente diferentes aspectos do ambiente, encontra na LSTM uma ferramenta poderosa. Sua capacidade de assimilar e processar múltiplas variáveis interdependentes permite a geração de previsões mais abrangentes e precisas, integrando informações de diversas fontes em um único modelo. Essa característica a credencia a ser utilizada em aplicações que exigem uma compreensão dinâmica complexa dos gases em ambientes monitorados.

As redes LSTM são particularmente adequadas para dados sequenciais, pois conseguem aprender e lembrar dependências temporais de longo prazo. Isso é crucial para séries temporais onde padrões importantes podem ocorrer em diferentes intervalos de tempo. Métodos estatísticos como ARIMA, SARIMA, ARIMAX e SARIMAX não conseguem produzir previsões de alvos múltiplos (ÖZDEN; BULUT, 2024).

5.1.1 Manuseio de Dados Não Lineares

As LSTMs, com sua capacidade de modelar relações não lineares, são particularmente vantajosas para séries temporais complexas, como dados meteorológicos, financeiros ou biológicos, que não se encaixam em padrões lineares simples. Essa característica permite que as LSTMs capturem nuances e interdependências intrínsecas presentes nesses tipos de dados.

Segundo (SHERSTINSKY, 2020), uma das limitações das redes neurais recorrentes tradicionais é o problema do *vanishing gradient*, onde gradientes diminutos impedem a rede de aprender dependências de longo prazo. As LSTMs foram especificamente desenvolvidas para mitigar esse problema, permitindo que a rede retenha informações relevantes ao longo de muitas etapas temporais. Incluir essa informação destaca a superioridade das LSTMs em relação às RNNs tradicionais.

5.1.2 Treinamento e Validação

O modelo foi treinado utilizando o otimizador Adam e a função de perda de erro quadrático médio (MSE). A escolha do otimizador Adam deve-se à sua eficiência computacional e à reduzida necessidade de ajuste de parâmetros. O Adam (*Adaptive Moment Estimation*) combina as vantagens dos métodos de otimização AdaGrad e RMSProp, permitindo um ajuste dinâmico das taxas de aprendizado para cada parâmetro. Ele calcula médias móveis dos gradientes e dos quadrados dos gradientes, o que proporciona uma convergência rápida e estável. Além disso, o Adam requer menos ajustes manuais dos hiperparâmetros, como a taxa de aprendizado, em comparação com outros otimizadores, tornando-o uma escolha robusta e eficaz para uma ampla gama de problemas de aprendizado de máquina.

A função de perda MSE foi selecionada por ser amplamente utilizada em problemas de regressão, devido à sua simplicidade e à capacidade de penalizar grandes erros. Para garantir que a rede neural não estivesse superajustada aos dados de treinamento e pudesse generalizar bem para novos dados, a validação do modelo foi realizada em um conjunto de dados separado.

5.2 Pré-processamento e Tratamento de Dados

O pré-processamento e tratamento de dados são etapas significativas em qualquer projeto de aprendizado de máquina, especialmente para modelos baseados em séries temporais, como as redes LSTM. A qualidade dos dados de entrada impacta diretamente a performance do modelo. A seguir, descrevemos os passos realizados para preparar os dados utilizados neste estudo.

Os dados foram inicialmente carregados a partir de um arquivo CSV contendo informações sobre a concentração de diferentes poluentes atmosféricos (HCHO, NO_2 e SO_2), umidade relativa (mínima e máxima), temperatura (mínima e máxima), velocidade do vento (mínima e máxima) e precipitação acumulada. Os dados foram organizados em um formato tabular, onde cada coluna representa uma variável de interesse e cada linha corresponde a um registro temporal. Essa organização tabular é fundamental para uma análise eficiente e precisa dos dados ambientais, permitindo insights valiosos para a gestão e mitigação da poluição atmosférica.

A Figura 9 apresenta a série temporal da concentração do formaldeído (HCHO) medida ao longo do período de 2018 a 2023. É possível observar que a concentração de HCHO apresenta 1 ou 2 picos durante o período analisado, com variações significativas em certos momentos, especialmente entre 2019 e 2023. Estes picos podem estar associados a eventos específicos, como períodos de maior emissão de poluentes, condições meteorológicas que favorecem a concentração de poluentes, ou outras atividades antropogênicas.

Os valores de concentração de HCHO permanecem em grande parte baixos, mas os picos sugerem que há momentos de maior emissão ou formação deste poluente. A análise detalhada destes picos pode fornecer subsídios valiosos para entender as causas subjacentes e implementar medidas corretivas adequadas.

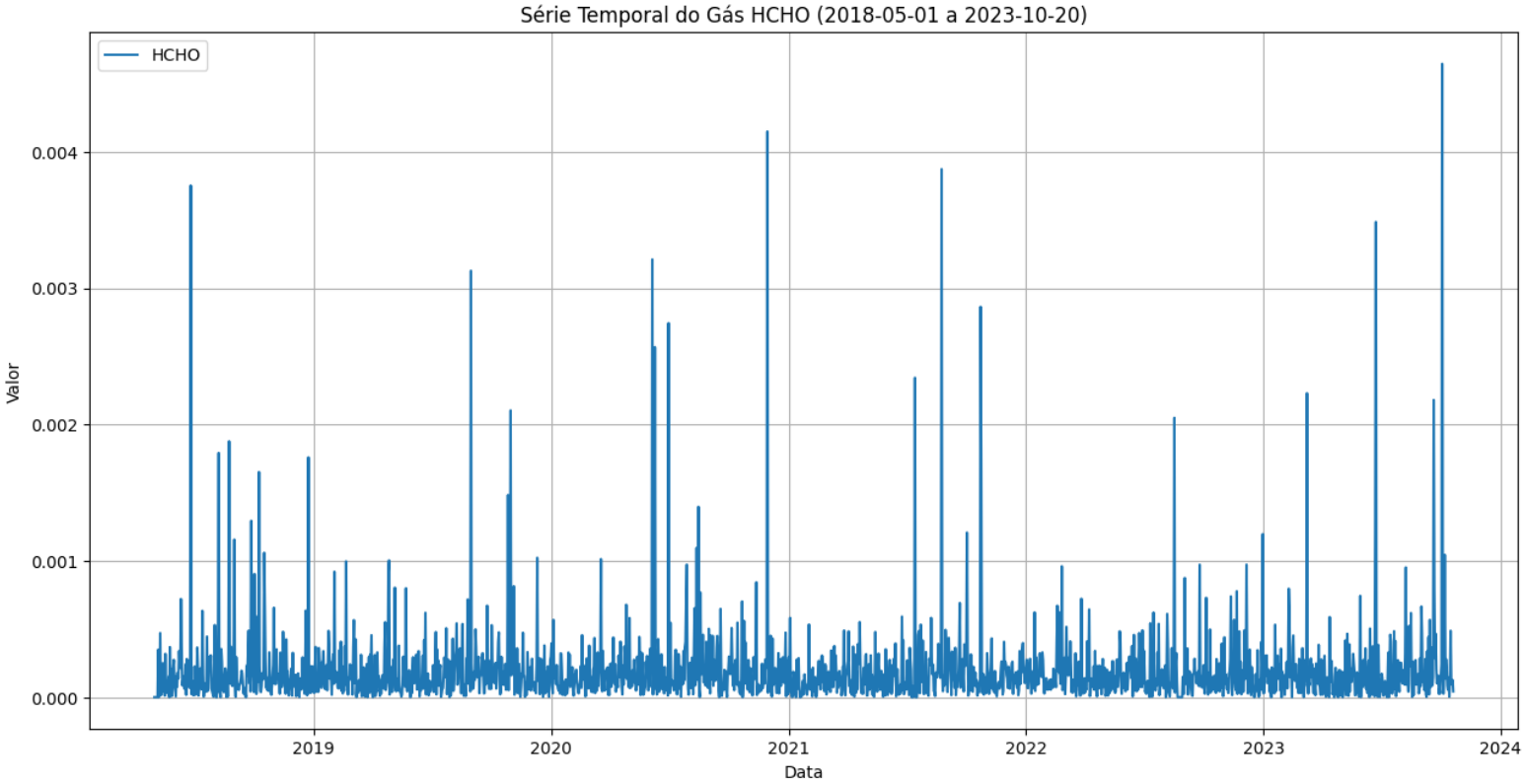
A Figura 10 apresenta a série temporal da concentração de SO_2 ao longo do período de 2018 a 2023. Similarmente à série de HCHO, a concentração deste gás também exibe picos esporádicos com variações significativas, principalmente entre 2020 e 2022.

Em contraste com o HCHO, o SO_2 apresenta um pico de concentração extremamente elevado, destoando significativamente dos demais valores observados. Esse pico sugere um evento de emissão ou formação de SO2 de grande magnitude, possivelmente relacionado a uma atividade industrial específica, um incidente ambiental ou condições meteorológicas particulares.

No geral, os valores de concentração de SO2 também se mantêm baixos na maior parte do tempo, indicando que as emissões ou formação desse poluente são geralmente controladas. No entanto, os picos observados, especialmente o de grande magnitude, evidenciam a necessidade de investigar as causas desses eventos e adotar medidas para prevenir ou mitigar suas ocorrências.

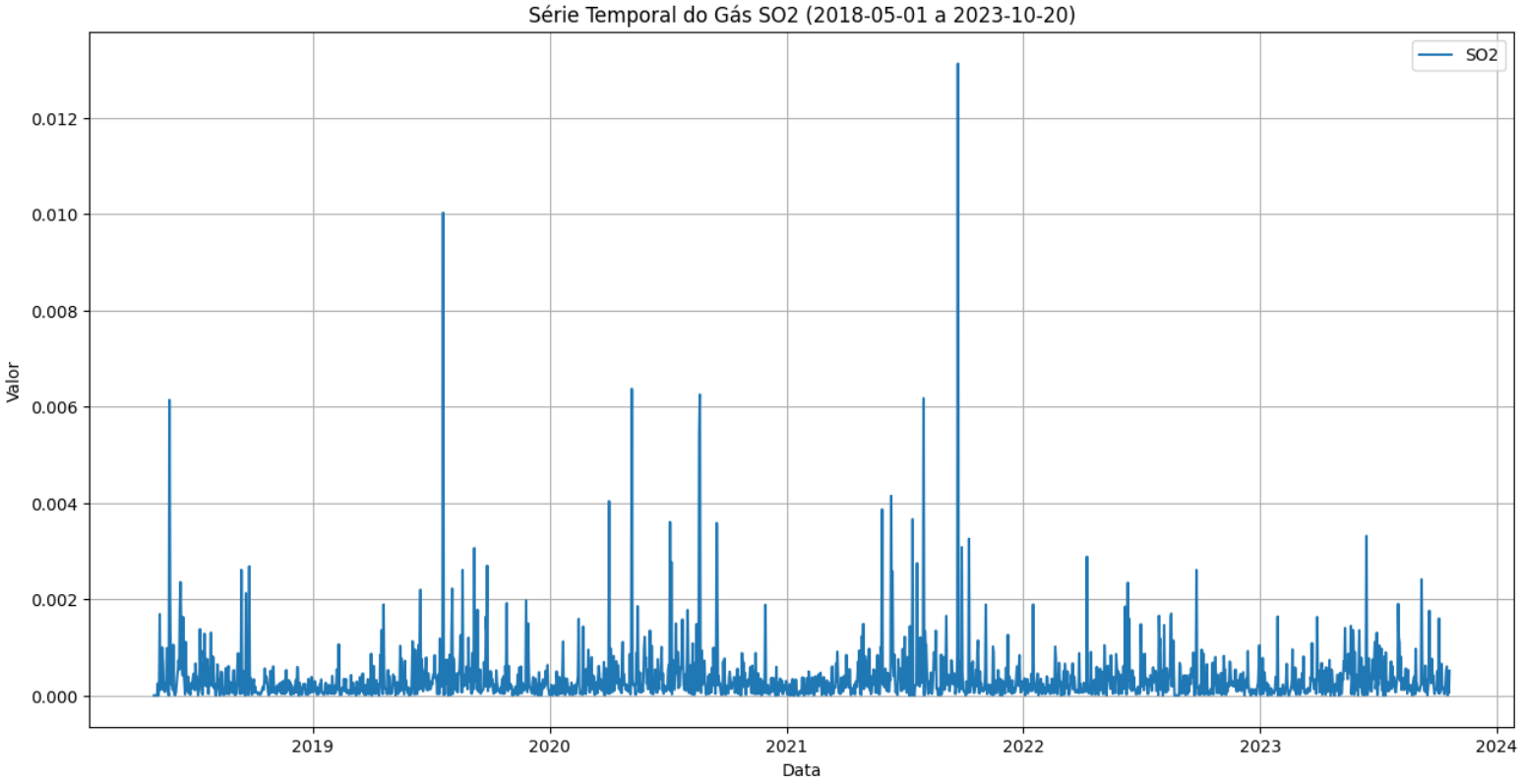
A série temporal da concentração de NO_2, apresentada na Figura 10, revela um comportamento distinto dos gases analisados anteriormente (HCHO e SO_2). Observa-se uma maior

Figura 9 – Serie temporal HCHO



Fonte: Autoria Própria - (2024).

Figura 10 – Serie temporal SO_2



Fonte: Autoria Própria - (2024).

frequência de picos ao longo de todo o período, com alguns picos mais pronunciados, especialmente no ano de 2022.

Diferentemente do HCHO e SO_2, que exibiam picos esporádicos em um fundo de baixas concentrações, o NO_2 demonstra uma flutuação mais constante, com picos e vales mais frequentes. Isso sugere que as fontes de emissão ou formação de NO_2 podem ser mais contínuas e menos sujeitas a eventos isolados de grande magnitude.

A análise da série temporal do NO2 também revela um possível padrão sazonal, com concentrações mais elevadas em determinados períodos do ano. Essa sazonalidade pode estar relacionada a fatores como variações no tráfego veicular, emissões industriais, condições meteorológicas e outros fatores que influenciam a formação e dispersão do NO_2.

5.3 Avaliação do Modelo

A avaliação do modelo de rede neural LSTM foi conduzida utilizando-se os dados reais e os dados de predição de concentração de gases e variáveis meteorológicas. A seguir, são apresentadas as tabelas com os dados utilizados e os gráficos comparativos para as diferentes concentrações de gases.

A Tabela 2 apresenta os dados reais de concentração dos gases HCHO, NO_2 e SO_2, além de variáveis meteorológicas como umidade relativa, temperatura, velocidade do vento e precipitação. Esses dados foram coletados entre os dias 21 e 31 de outubro de 2023.

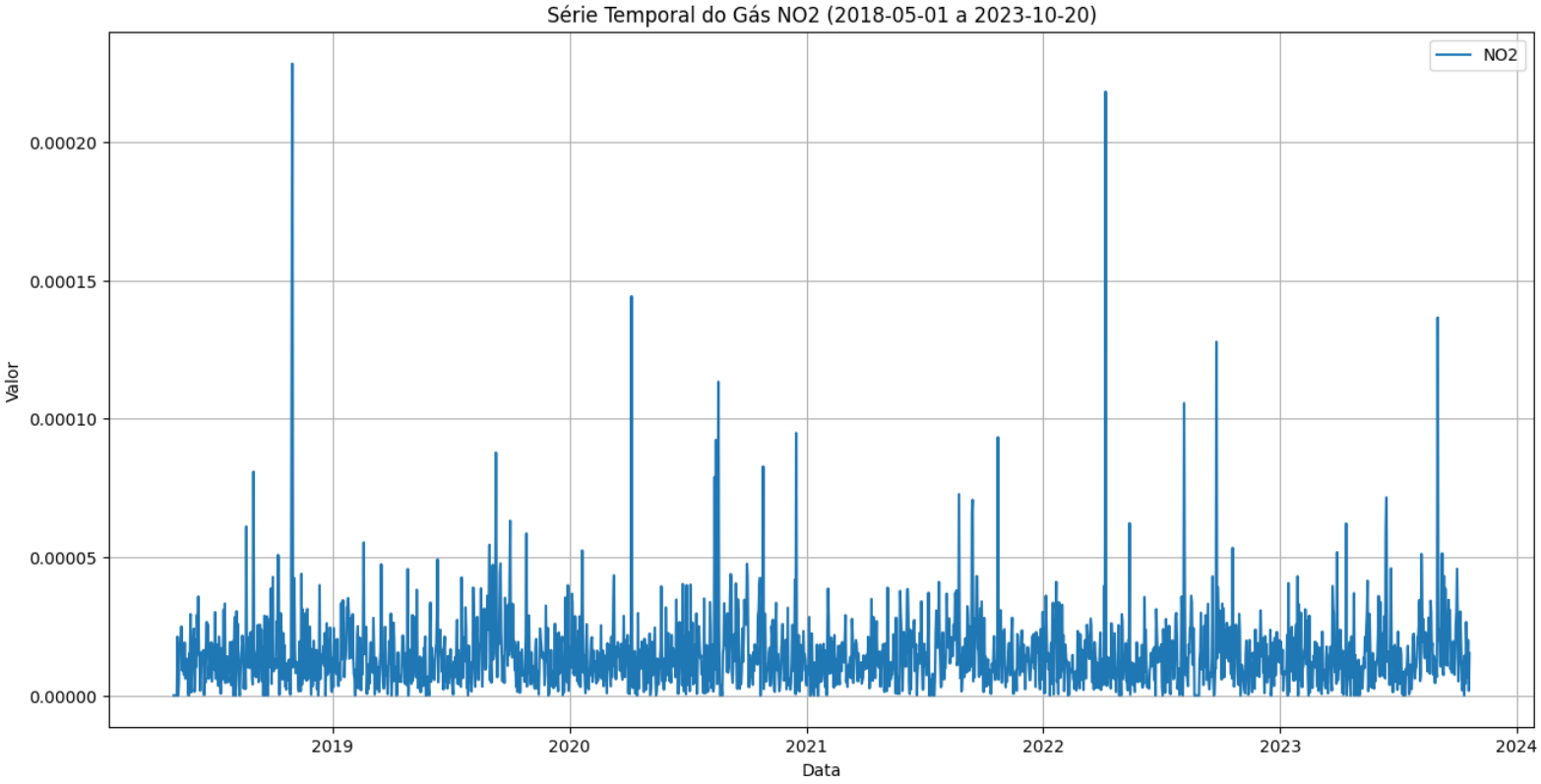
Tabela 2 – Dados de Concentração Reais de Gases mol/m^2 e Variáveis Meteorológicas

Data	HCHO	NO_2	SO_2	Vel.Vento Min.	Vel.Vento Max.	Precip. Total
2023-10-21	0,000086	-	0,000206	0,00	3,40	0,20
2023-10-22	0,000104	0,000067	0,000029	0,00	4,90	0,00
2023-10-23	0,000468	0,000022	0,000155	0,00	9,10	9,20
2023-10-24	0,000197	0,000007	0,000152	0,00	3,80	0,80
2023-10-25	0,000153	0,000016	0,000066	0,00	5,30	0,20
2023-10-26	0,000091	0,000015	0,000371	0,00	6,10	0,00
2023-10-27	0,000156	0,000034	0,000169	0,00	6,50	26,80
2023-10-28	0,000661	0,000021	0,000139	0,00	10,30	181,60
2023-10-29	0,000031	0,000002	0,000110	0,00	6,50	1,40
2023-10-30	0,000118	0,000019	0,000115	0,00	11,40	41,20
2023-10-31	0,000065	0,000014	0,000056	0,00	3,80	2,20

Fonte: Autoria Própria (2024).

A Tabela 3 contém os dados de predição gerados pelo modelo LSTM. Esses dados representam as previsões feitas para as mesmas datas e variáveis apresentadas na Tabela 1, permitindo uma comparação direta entre os valores reais e os valores preditos pelo modelo.

Figura 11 – Serie temporal NO_2



Fonte: Autoria Própria (2024).

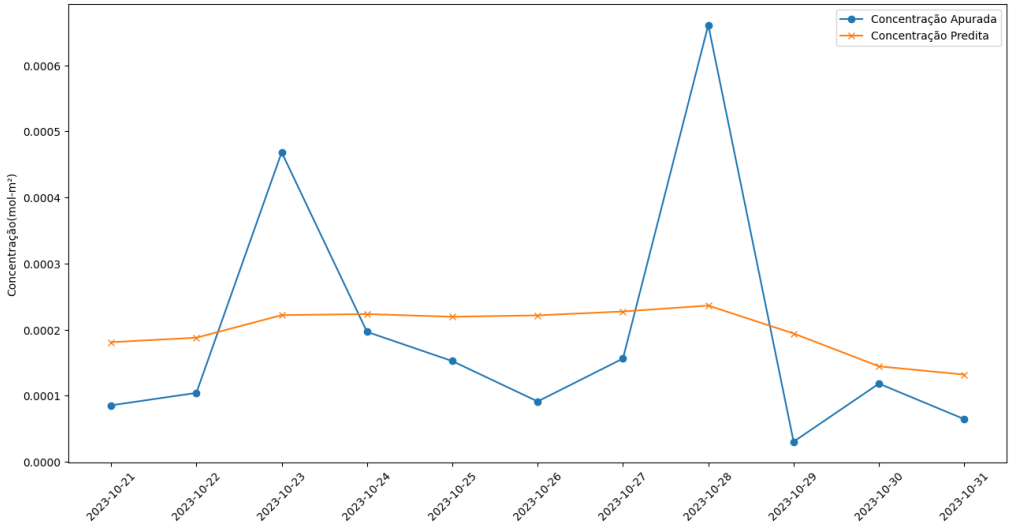
Tabela 3 – Dados de Predição de Concentração de Gases (mol/m^2) e Variáveis Meteorológicas

Data	HCHO	NO_2	SO_2	Vel.Vento Min.	Vel.Vento Max.	Precip. Total
2023-10-21	0.000181	0.000011	0.000238	0.00	1.22	4.55
2023-10-22	0.000188	0.000011	0.000241	0.00	0.86	3.81
2023-10-23	0.000222	0.000014	0.000363	0.00	0.42	3.57
2023-10-24	0.000224	0.000013	0.000357	0.00	0.47	2.86
2023-10-25	0.000220	0.000013	0.000337	0.00	0.34	2.34
2023-10-26	0.000222	0.000013	0.000330	0.00	0.54	2.38
2023-10-27	0.000228	0.000013	0.000328	0.00	0.81	2.17
2023-10-28	0.000237	0.000013	0.000302	0.00	1.49	3.45
2023-10-29	0.000194	0.000006	0.000000	0.00	5.96	12.41
2023-10-30	0.000145	0.000006	0.000030	0.00	3.59	7.45
2023-10-31	0.000132	0.000007	0.000060	0.00	3.67	9.14

Fonte: Autoria Própria (2024).

A Figura 12 mostra a comparação entre os valores reais e os valores preditos de concentração de HCHO ao longo do período analisado. Observa-se que as previsões do modelo apresentam uma tendência similar aos dados reais, embora com algumas discrepâncias em determinados pontos.

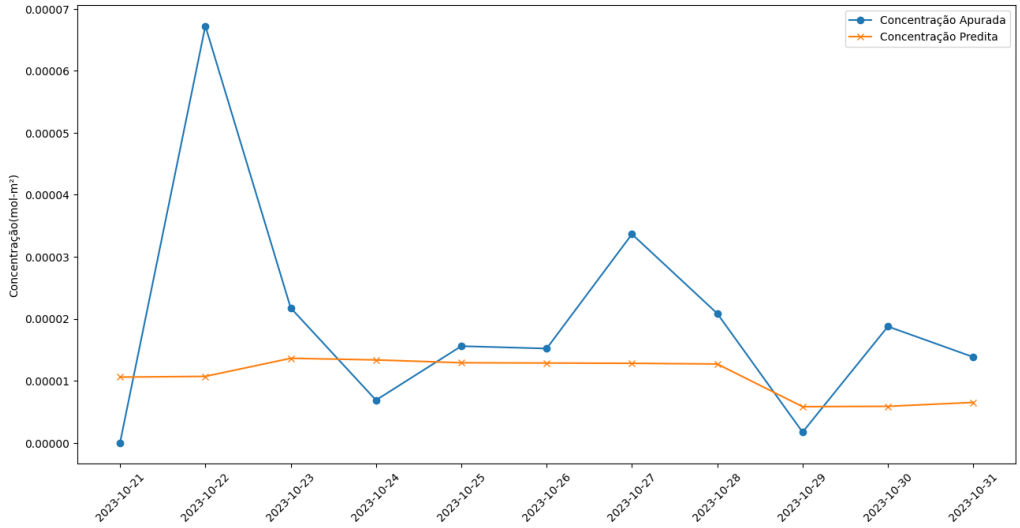
Figura 12 – Comparação entre os valores reais e preditos de concentração de HCHO(mol/m^2)



Fonte: Autoria Própria (2024).

Na Figura 13, é apresentada a comparação entre os valores reais e preditos de concentração de NO2. Assim como na análise de HCHO, as previsões do modelo seguem a tendência dos dados reais, mas com variações que podem indicar áreas de melhoria para o modelo.

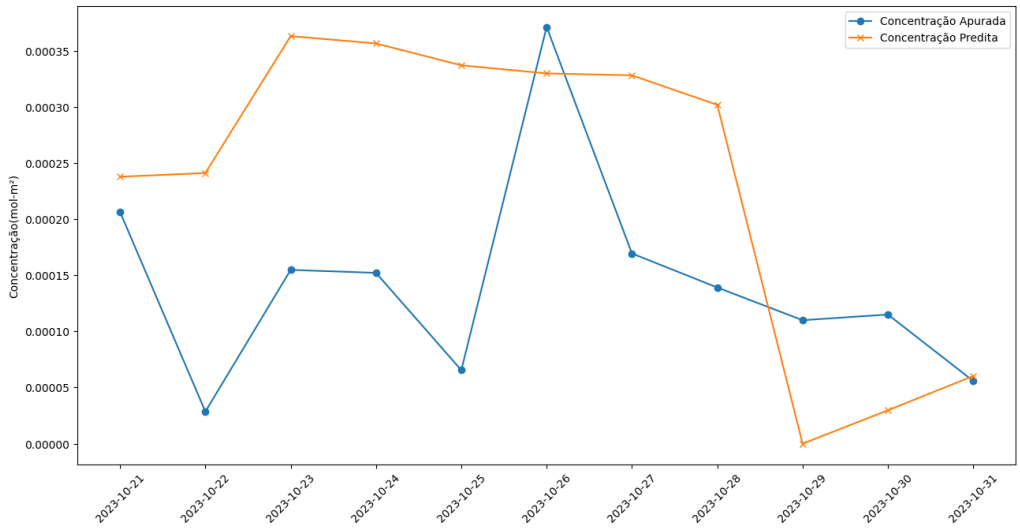
Figura 13 – Comparação entre os valores reais e preditos de concentração de NO_2 (mol/m^2)



Fonte: Autoria Própria (2024).

A Figura 14 ilustra a comparação entre os valores reais e preditos de concentração de SO_2. Mais uma vez nota-se que o modelo consegue capturar a tendência geral dos dados reais, porém, as discrepâncias sugerem a necessidade de ajustes no modelo para melhorar a precisão das previsões.

Figura 14 – Comparação entre os valores reais e preditos de concentração de SO_2(mol/m^2)



Fonte: Autoria Própria (2024).

A análise gráfica das tabelas acima demonstra que o modelo LSTM é capaz de prever a tendencia das concentrações dos gases HCHO, NO_2 e SO_2 com um grau de precisão razoável. No entanto, as diferenças observadas entre os valores reais e preditos indicam que

o modelo ainda possui limitações. Essas discrepâncias podem ser atribuídas a vários fatores, como a complexidade dos dados, a necessidade de maior quantidade de dados de treinamento ou ajustes nos hiperparâmetros do modelo.

Para quantificar a precisão do modelo, foram calculadas métricas como o erro quadrático médio (MSE), a raiz do erro quadrático médio (RMSE) e o erro absoluto médio (MAE) para cada um dos gases analisados, de acordo com a tabela 4. Essas métricas fornecem uma medida quantitativa da diferença entre os valores reais e preditos, ajudando a identificar áreas específicas para aprimoramento do modelo.

Tabela 4 – Métricas de Erro para Predições de Concentração de Gases

Gás	MSE	RMSE	MAE
HCHO	$2{,}88 \times 10^{-8}$	$1{,}70 \times 10^{-4}$	$1{,}28 \times 10^{-4}$
NO_2	$3{,}79 \times 10^{-10}$	$1{,}95 \times 10^{-5}$	$1{,}27 \times 10^{-5}$
SO_2	$2{,}53 \times 10^{-8}$	$1{,}59 \times 10^{-4}$	$1{,}36 \times 10^{-4}$

Fonte: Autoria Própria (2024).

5.3.1 Interpretação dos Resultados

- HCHO - O valor do erro quadrático médio (MSE) para o HCHO é $2{,}8757 \times 10^{-8}$. Esse valor representa a média dos quadrados das diferenças entre os valores preditos e os valores reais. Quanto menor o MSE, melhor é a performance do modelo. No caso do HCHO, o MSE é relativamente baixo, indicando que as previsões do modelo estão próximas dos valores reais. O RMSE, que é a raiz quadrada do MSE, é $1{,}6958 \times 10^{-4}$. Esse valor fornece uma medida da magnitude do erro e é expresso nas mesmas unidades dos dados originais. O RMSE baixo sugere que o modelo é bastante preciso nas previsões de HCHO. O MAE para o HCHO é $1{,}2751 \times 10^{-4}$. Esse valor representa a média das diferenças absolutas entre os valores preditos e os valores reais, sendo uma medida direta do erro médio. Novamente, o baixo valor do MAE indica uma boa precisão do modelo.

- NO_2 - O valor do MSE para o NO_2 é $3{,}7885 \times 10^{-10}$, extremamente baixo, indicando uma excelente performance do modelo. O RMSE é $1{,}9464 \times 10^{-5}$, reforçando que as previsões do modelo estão muito próximas dos valores reais. O MAE para o NO_2 é $1{,}2735 \times 10^{-5}$. Esse valor mostra que, em média, as previsões do modelo diferem dos valores reais por uma pequena margem, confirmando a alta precisão do modelo.

- SO_2 - O MSE para o SO_2 é $2{,}5254 \times 10^{-8}$, um valor também baixo, sugerindo uma boa performance do modelo. O RMSE é $1{,}5892 \times 10^{-4}$, indicando que o modelo consegue prever os valores de SO_2 com boa precisão. O MAE para o SO_2 é $1{,}3551 \times 10^{-4}$. Assim como as outras métricas, esse valor baixo sugere que o modelo está fazendo previsões precisas.

Os resultados das métricas MSE, RMSE e MAE para os gases HCHO, NO_2 e SO_2 indicam que o modelo de rede neural LSTM possui um bom desempenho na previsão das concentrações desses gases. Em particular, o modelo mostrou uma precisão notável na predição de NO2, com os valores de MSE, RMSE e MAE sendo baixos.

As previsões de HCHO e SO_2 também apresentaram boa precisão, embora com uma ligeira variação em comparação com NO_2. Essas variações podem ser atribuídas à natureza dos dados ou a possíveis melhorias que ainda podem ser feitas no modelo, como o ajuste dos hiperparâmetros ou a inclusão de mais dados de treinamento.

A avaliação do modelo de rede neural LSTM demonstrou que ter uma boa eficacia na previsão das concentrações dos gases, uma vez que todas as métricas indicam um bom grau de precisão. No entanto, para melhorar que se possa obter mais a performance do modelo, recomenda-se explorar ajustes adicionais nos parâmetros do modelo, aumentar o conjunto de dados de treinamento e considerar técnicas avançadas de regularização e validação.

Os resultados promissores obtidos nesta análise destacam o potencial das redes neurais LSTM na previsão de concentrações de gases, abrindo caminho para futuras pesquisas e aplicações práticas no monitoramento e controle da qualidade do ar.

5.4 Aplicações Práticas e Impactos

A precisão e a capacidade de generalização de modelos preditivos têm um papel crucial em diversas aplicações práticas, afetando diretamente a tomada de decisões em diferentes setores. Com base nos resultados obtidos, é possível explorar como as previsões bem-sucedidas e as limitações do modelo podem influenciar essas áreas.

Monitoramento e Controle Ambiental: A utilização de modelos preditivos como o LSTM para prever concentrações de gases é fundamental no monitoramento e controle ambiental. A capacidade do modelo de prever com precisão os níveis de HCHO, NO_2 e SO_2 permite que autoridades ambientais respondam rapidamente a níveis perigosos de poluição do ar. Isso é crucial para a implementação de medidas preventivas e corretivas, como a restrição de atividades industriais em dias de alta poluição ou a emissão de alertas à população.

Saúde Pública: A precisão das previsões de poluentes atmosféricos impacta diretamente a saúde pública. Altos níveis de HCHO, NO_2 e SO_2 estão associados a uma série de problemas de saúde, incluindo doenças respiratórias e cardiovasculares. Com previsões precisas, é possível alertar a população vulnerável (como crianças, idosos e pessoas com condições respiratórias) para que tomem precauções. Além disso, autoridades de saúde podem planejar campanhas de conscientização e intervenções médicas mais promissoras.

Agricultura e Gestão de Recursos Hídricos: A qualidade do ar também afeta a agricultura e a gestão de recursos hídricos. Poluentes como NO_2 podem impactar a saúde das plantas, reduzindo a produtividade agrícola. Modelos preditivos precisos permitem que agricultores ajustem suas práticas de cultivo e irrigação para mitigar os impactos da poluição. Na gestão

de recursos hídricos, a previsão de precipitação (incluída como uma variável meteorológica no modelo) pode ajudar a planejar a alocação de água e a prevenir inundações ou secas.

Planejamento Urbano e Infraestrutura: O planejamento urbano e a infraestrutura podem se beneficiar enormemente de previsões precisas de poluentes e variáveis meteorológicas. Tomar decisões informadas sobre a localização de zonas industriais, áreas residenciais e espaços verdes pode ajudar a minimizar a exposição da população à poluição. Além disso, a previsão de eventos climáticos extremos pode auxiliar na preparação e resposta a desastres, melhorando a resiliência das cidades.

Desenvolvimento de Políticas Publicas: A elaboração de políticas públicas sólidas depende de dados precisos e previsões confiáveis. As previsões de poluentes atmosféricos fornecidas por modelos LSTM podem informar a criação de regulamentos ambientais mais promissoras. Políticas para reduzir as emissões de poluentes podem ser desenvolvidas com base em previsões de alta precisão, garantindo que os esforços para melhorar a qualidade do ar sejam direcionados e promissoras. Além disso, essas previsões podem apoiar a transparência e a comunicação com o público, aumentando a confiança nas ações governamentais.

As previsões precisas fornecidas pelo modelo LSTM têm o potencial de impactar positivamente várias áreas, desde a saúde pública até o desenvolvimento de políticas públicas. No entanto, é importante reconhecer e abordar as limitações do modelo para maximizar sua eficácia e aplicabilidade. A continuidade na pesquisa e no desenvolvimento de modelos preditivos, juntamente com a integração de novas tecnologias e dados, será essencial para enfrentar os desafios ambientais e sociais do futuro.

6 DISCUSSÕES

6.1 Limitações do Estudo

Um dos principais desafios enfrentados neste estudo foi a ausência de uma estação de coleta no centro da cidade, o que restringiu a abrangência e a precisão dos dados disponíveis para análise. A falta de infraestrutura adequada para a captação de dados dos gases estudados, como HCHO, NO_2 e SO_2, compromete a qualidade das informações, afetando negativamente a precisão das previsões do modelo. Além disso, a quantidade e a qualidade dos dados disponíveis são insuficientes para uma análise mais robusta, limitando a confiabilidade dos resultados obtidos.

Outro aspecto crucial é a habilidade de estender o modelo a novos dados. O modelo pode não se ajustar adequadamente a novos ambientes ou comportamentos que não foram identificados durante o treinamento, levando a previsões menos acuradas. A incorporação de variáveis importantes que não foram levadas em conta no modelo vigente também constitui uma restrição considerável. Fatores interrelacionados que afetam as concentrações de gases podem não ter sido considerados, restringindo a acurácia e a utilidade das previsões.

As limitações de modelagem, como a seleção de hiperparâmetros e a arquitetura do modelo, também desempenham um papel crucial no desempenho preditivo. Escolhas podem resultar em modelos que não capturam adequadamente as complexidades dos dados, levando a previsões imprecisas. A interpretação dos resultados pode ser desafiadora devido à complexidade do modelo, tornando difícil entender como diferentes variáveis influenciam as previsões.

Por fim, a implementação de modelos preditivos avançados requer recursos computacionais significativos e expertise em ciência de dados. Organizações que não possuem essa infraestrutura ou conhecimentos especializados podem encontrar dificuldades na aplicação e manutenção desses modelos, limitando seu uso prático.

6.2 Desafios Futuros

Para superar essas limitações, é essencial melhorar a qualidade e a disponibilidade dos dados. A instalação de estações de coleta no centro da cidade e o aprimoramento da infraestrutura para a captação de dados dos gases estudados são passos cruciais. Garantir um volume maior e mais preciso de dados permitirá um treinamento e validação mais robustos dos modelos preditivos.

Desenvolver técnicas que aprimorem a generalização do modelo para novos dados e diferentes condições ambientais é um desafio contínuo. Isso inclui a exploração de novas abordagens de modelagem e a otimização das existentes para garantir que o modelo se adapte bem a mudanças nas condições.

A integração de variáveis adicionais que influenciem as concentrações de gases pode aumentar significativamente a precisão das previsões. Considerar múltiplos fatores interdependentes permitirá uma visão mais abrangente dos processos que afetam a qualidade do ar.

Avanços em técnicas de modelagem são essenciais para melhorar o desempenho preditivo. A exploração de novas metodologias e a contínua otimização dos modelos existentes são fundamentais para capturar as complexidades dos dados e fornecer previsões mais precisas.

Tornar os modelos mais interpretáveis e transparentes também é importantíssimo. Desenvolver métodos que facilitem a compreensão de como diferentes variáveis afetam as previsões ajudará na aceitação e confiança nos resultados, além de facilitar a tomada de decisões informadas.

Finalmente, investir em infraestrutura computacional robusta e capacitação de profissionais em ciência de dados é fundamental. Isso permitirá a implementação eficaz e a manutenção de modelos preditivos avançados, garantindo que os benefícios dessas tecnologias sejam plenamente realizados.

Ademais as limitações atuais e enfrentar os desafios futuros permitirá avanços significativos no uso de modelos preditivos para monitoramento ambiental, saúde pública e outras áreas críticas. A continuidade na pesquisa e desenvolvimento, aliada a uma melhoria na qualidade dos dados e na infraestrutura disponível, será essencial para maximizar os benefícios desses modelos.

Os desafios futuros no desenvolvimento e aplicação de modelos preditivos são multifacetados e complexos. Abordar esses desafios requer um esforço colaborativo e contínuo para melhorar a qualidade dos dados, a capacidade de generalização, a inclusão de variáveis relevantes, o avanço das técnicas de modelagem, a interpretação e transparência, a disponibilização de recursos e a consideração de questões éticas. Superar esses desafios permitirá que os modelos preditivos atinjam seu pleno potencial e ofereçam benefícios significativos em diversas áreas de aplicação.

7 CONCLUSÃO

7.1 Resultados e Contribuições

Os resultados mais significativos deste estudo incluem a precisão das previsões de concentrações de NO_2, SO_2 e HCHO, demonstrando o potencial dos modelos preditivos em aplicações ambientais e de saúde pública. A análise revelou tanto sucessos quanto áreas de melhoria, fornecendo direções claras para futuras pesquisas. A metodologia empregada e a abordagem de modelagem utilizadas provaram ser promissores, mas também destacaram limitações que devem ser abordadas. As contribuições deste estudo aumentam o entendimento sobre a eficácia dos modelos preditivos e promovem seu uso responsável e informado em várias áreas, beneficiando a sociedade em geral.

As implicações práticas dos modelos preditivos desenvolvidos são vastas e diversas. No monitoramento ambiental e da qualidade do ar, eles podem ajudar a identificar áreas de alto risco e permitir a implementação de medidas preventivas. Na saúde pública, as previsões precisas de poluentes atmosféricos podem informar políticas de saúde e campanhas de sensibilização, ajudando a mitigar os impactos negativos da poluição. No planejamento urbano e gestão de infraestrutura, esses modelos podem auxiliar no desenvolvimento de estratégias para reduzir a poluição e melhorar a qualidade de vida urbana. Na agricultura e gestão de recursos naturais, as previsões de variáveis ambientais podem otimizar o uso de recursos e aumentar a produtividade. No desenvolvimento de políticas, os modelos preditivos fornecem uma base sólida para a formulação de políticas informadas e baseadas em evidências, promovendo um ambiente mais saudável e sustentável.

No entanto, a ausência de uma estação meteorológica para coleta de MP10 no centro da cidade representa uma limitação significativa. Investimentos em infraestrutura de coleta de dados e esforços contínuos para melhorar a precisão e a abrangência dos modelos são essenciais para maximizar os benefícios dessas tecnologias.

Há várias oportunidades futuras para expandir e melhorar este trabalho. A expansão e melhoria da coleta de dados, incluindo a instalação de novas estações de monitoramento, são cruciais. A integração de dados multidimensionais, como informações meteorológicas, geográficas e sociais, pode melhorar significativamente a precisão dos modelos. O desenvolvimento de modelos híbridos que combinem diferentes abordagens de modelagem pode oferecer soluções mais robustas. A exploração de técnicas avançadas de machine learning, como deep learning e redes neurais, pode levar a avanços significativos. Aplicações em tempo real dos modelos preditivos podem fornecer alertas rápidos e permitir uma resposta imediata a condições adversas. A participação comunitária e cidadã pode ser promovida para aumentar a conscientização e a colaboração no monitoramento ambiental. Finalmente, o desenvolvimento de políticas baseadas em evidências, utilizando dados precisos e previsões confiáveis, pode resultar em ações mais promissores e sustentáveis.

O estudo sobre modelos preditivos revela um panorama complexo e dinâmico para a previsão de condições ambientais e variáveis atmosféricas. Os resultados destacam tanto o potencial significativo quanto as limitações desses modelos, oferecendo uma visão valiosa sobre suas aplicações práticas e desafios. Os principais resultados e contribuições incluem a precisão das previsões, a identificação de limitações e a orientação para futuras pesquisas. As implicações práticas são vastas, abrangendo monitoramento ambiental, saúde pública, planejamento urbano, agricultura e desenvolvimento de políticas.

Os desafios identificados, como a necessidade de melhor infraestrutura e a integração de mais variáveis, apresentam oportunidades futuras para aprimoramentos. Investimentos estratégicos em infraestrutura, avanços em técnicas de modelagem e a colaboração ativa da comunidade são essenciais para alcançar benefícios significativos. A contínua inovação e melhoria dos modelos preditivos visam um futuro mais informado e sustentável, com impactos positivos na saúde pública, no planejamento urbano e na sustentabilidade ambiental.

REFERÊNCIAS

ABDULKADIR, S. J. *et al.* Long short term memory recurrent network for standard and poorâ€™s 500 index modelling. **International Journal of Engineering Technology**, v. 7, n. 4.15, p. 25–29, Oct. 2018. Disponível em: https://www.sciencepubco.com/index.php/ijet/article/view/21365.

ABIODUN, O. I. *et al.* State-of-the-art in artificial neural network applications: A survey. **Heliyon**, Elsevier, v. 5, n. 11, p. e01338, 2019.

ADAMS, M. *et al.* Comparison of linear regression and neural network models forecasting tourist arrivals to turkey. **Eurasian Journal of Science & Engineering**, 2015.

AMÂNCIO, C. T.; NASCIMENTO, L. F. C. Asma e poluentes ambientais: um estudo de séries temporais. **Revista da Associação Médica Brasileira**, SciELO Brasil, v. 58, p. 302–307, 2012.

AMBIENTE, C. C. N. D. M. Resolução nº 491, de 19 de novembro de 2018. dispõe sobre padrões de qualidade do ar. **Diário Oficial d'a União**, 2018.

AVELINO, W. de O. *et al.* Implementação de redes neurais por pulsos a partir de sinapses memristivas. Universidade Federal de Minas Gerais, 2022.

BRITO, J.; BERNARDO, A.; GONÇALVES, L. L. Atmospheric pollution and mortality in portugal: Quantitative assessment of the environmental burden of disease using the airq+ model. **Science of the total environment**, Elsevier, v. 815, p. 152964, 2022.

CANÇADO, J. E. D. *et al.* Repercussões clínicas da exposição à poluição atmosférica. **Jornal Brasileiro de Pneumologia**, SciELO Brasil, v. 32, p. S5–S11, 2006.

CHAVES, L. E. *et al.* Modelos computacionais fuzzy e neuro-fuzzy para avaliarem os efeitos da poluição do ar. Universidade Estadual Paulista (UNESP), 2013.

COSTA, M. A. *et al.* Efeitos das partículas totais em suspensão (pts) na saúde da população dos bairros de benfica, bonsucesso, ramos e manguinhos-rio de janeiro/rj. **Hygeia Revista Brasileira de Geografia Medica e de saude**, Associação Nacional de Pesquisa e Pós-Graduação em Geografia, Grupo de . . . , v. 5, n. 9, 2009.

COSTA, M. L.; SOUZA, R. A. Desenvolvimento regional e infraestrutura urbana em francisco beltrão-pr. **Revista Paranaense de Desenvolvimento**, Instituto Paranaense de Desenvolvimento Econômico e Social, v. 41, n. 2, p. 45–62, 2020.

FARIAS, A.; MENDONCA, F. Riscos socioambientais de inundação urbana sob a perspectiva do sistema ambiental urbano. **Sociedade e Natureza**, Editora da Universidade Federal de Uberlândia - EDUFU, v. 34, p. e63717, 2022. ISSN 1982-4513. Disponível em: https://doi.org/10.14393/SN-v34-2022-63717.

GUO, Z.-D. *et al.* Aerosol and surface distribution of severe acute respiratory syndrome coronavirus 2 in hospital wards, wuhan, china, 2020. **Emerging infectious diseases**, Centers for Disease Control and Prevention, v. 26, n. 7, p. 1586, 2020.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Frota de Veículos - Francisco Beltrão - PR.** 2022. Dados do Ministério da Infraestrutura, SENATRAN - Secretaria Nacional de Trânsito. Disponível em: https://cidades.ibge.gov.br/brasil/pr/francisco-beltrao/pesquisa/22/28120?tipo=grafico&indicador=28120.

IBGE, I. B. de Geografia e E. **Censo Demográfico 2023**. 2023. Acesso em [2024-07-10]. Disponível em: https://cidades.ibge.gov.br/brasil/pr/francisco-beltrao/panorama.

JAGTAP, A. D.; KAWAGUCHI, K.; KARNIADAKIS, G. E. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. **Journal of Computational Physics**, Elsevier, v. 404, p. 109136, 2020.

KAMALOV, F.; SULIEMAN, H. Time series signal recovery methods: comparative study. p. 1–5, 2021.

KUMAR, A.; JAIN, V. K.; KUMAR, R. Arima forecasting of ambient air pollutants (o3, no, no2 and co). **Stochastic Environmental Research and Risk Assessment**, Springer, v. 30, n. 2, p. 639–652, 2016.

LIU, H. *et al.* Deep learning for air quality forecasting: A survey. **Journal of Cleaner Production**, Elsevier, v. 217, p. 135–144, 2019.

MACEDO, R. L. Qualidade do ar em campo grande–ms: estudo das emissões por fontes móveis e sua dispersão. 2012.

PESTANA, P. R. d. S. Relação entre poluição do ar e internações hospitalares por doenças respiratórias e cardiovascular em uma cidade do interior do estado de são paulo. Universidade Estadual Paulista (UNESP), 2015.

RAMOS, J. P. S. Redes neurais artificiais na classificação de frutos: cenário bidimensional. **Ciência e Agrotecnologia**, Editora da Universidade Federal de Lavras, v. 27, n. 2, p. 356–362, Mar 2003. ISSN 1413-7054. Disponível em: https://doi.org/10.1590/S1413-70542003000200015.

RIBEIRO, A. D. M. Implementação em fpga de algoritmos para o controle de motores trifásicos baseados em redes neurais artificiais com função de ativação projetada usando algoritmos genéticos. 2020.

SANTOS, M. R. F. d. Padrões da qualidade do ar durante a paralisação nacional dos caminhoneiros no ano de 2018, no município de canoas-rs. 2021.

SARKER, I. H. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. **SN Computer Science**, Springer, v. 2, n. 6, p. 1–20, 2021.

SHERSTINSKY, A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. **Physica D: Nonlinear Phenomena**, Elsevier, v. 404, p. 132306, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167278919305974.

SILVA, L. F. d. **Avaliação da evolução das concentrações dos principais poluentes atmosféricos monitorados no município de Curitiba-PR**. 2015. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2015.

SILVEIRA, T. S. **Qualidade do ar e concentração de materiais particulados nos principais centros urbanos da América do Sul: análise multidisciplinar a partir de relatórios oficiais considerando aspectos de geografia, direito e saúde coletiva**. 2022. Tese (Doutorado) — Universidade de São Paulo, 2022.

SUN, S. *et al.* A survey of optimization methods for deep learning. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 31, n. 11, p. 5145–5173, 2020.

TAETS, L. A. R. Análise da participação do gás natural na matriz energética e a aplicação desse combustível no processo de pelotização de minério de ferro. 2014.

TANG, Y.; WANG, S. Air quality prediction using machine learning: Current trend and future directions. **Environmental Pollution**, Elsevier, v. 277, p. 116727, 2021.

VALDAMBRINI, N. M.; RIBEIRO, F. N. D. Avaliação das ultrapassagens dos padrões de ozônio troposférico no estado de são paulo de 2014 a 2019. **Revista Brasileira de Meteorologia**, Sociedade Brasileira de Meteorologia, v. 36, n. 4, p. 735–747, Oct 2021. ISSN 0102-7786. Disponível em: https://doi.org/10.1590/0102-7786360046.

WANG, C. *et al.* A lstm approach with word embedding for short text classification. **Soft Computing**, Springer, v. 24, n. 14, p. 10233–10243, 2020.

WANG, Y.; DI, B.; ZHANG, W. Application of deep learning in the prediction of air quality. **Journal of Cleaner Production**, Elsevier, v. 248, p. 119214, 2020.

WHO. **Air quality guidelines for Europe**. 2. ed. Copenhagen: WHO Regional Publications, European Series, No. 91, 2000. 273 p. ISBN 9789289013581. Disponível em: https://iris.who.int/bitstream/handle/10665/107335/9789289013581-eng.pdf?sequence=1.

Yuan, X. *et al.* Monthly runoff forecasting based on LSTM-ALO model. **Stochastic Environmental Research and Risk Assessment**, v. 32, n. 8, p. 2199–2212, ago. 2018.

ZHAO, Z.; ZHANG, Y.; XU, X. A comparison of air quality prediction models based on lstm and arima. **Atmospheric Environment**, Elsevier, v. 214, p. 116884, 2019.

ZHOU, T. *et al.* Lstm network: A deep learning approach for short-term traffic forecasting. **IET Intelligent Transport Systems**, IET, v. 14, n. 9, p. 1064–1070, 2020.

ÖZDEN, C.; BULUT, M. Spectral temporal graph neural network for multivariate agricultural price forecasting. **Ciência Rural**, Universidade Federal de Santa Maria, v. 54, n. 1, p. e20220677, 2024. ISSN 0103-8478. Disponível em: https://doi.org/10.1590/0103-8478cr20220677.

ANEXOS

ANEXO A – Códigos Fonte

Listagem 1 – Importação de Bibliotecas

```
import sys
import pandas as pd
import numpy as np
import tensorflow as tf
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
```

Fonte: Autoria própria (2024).

Listagem 2 – Carregamento e Preparação dos Dados

```
# Carregar os dados do arquivo CSV
df = pd.read_csv('seus_dados.csv', sep=';')
df['data'] = pd.to_datetime(df['data'])
df.set_index('data', inplace=True)
```

Fonte: Autoria própria (2024).

Listagem 3 – Normalização dos Dados

```
# Normalizar os dados
scaler = MinMaxScaler()
df_normalized = pd.DataFrame(scaler.fit_transform(df),
columns=df.columns, index=df.index)
```

Fonte: Autoria própria (2024).

Listagem 4 – Criação de Sequências

```
def create_sequences(data, seq_length):
    X = []
    y = []
    for i in range(len(data) - seq_length):
        X.append(data.iloc[i:i+seq_length].values)
        y.append(data.iloc[i+seq_length].values)
        # Corrigido para capturar o próximo valor corretamente
    return np.array(X), np.array(y)
```

Fonte: Autoria própria (2024).

Listagem 5 – Comprimento da Sequência

```
seq_length = 10
```

Fonte: Autoria própria (2024).

Listagem 6 – Divisão dos Dados

```
train_data = df_normalized.loc['2018-01-01':'2022-12-31']
validation_data = df_normalized.loc['2023-01-01':'2023-10-20']
```

Fonte: Autoria própria (2024).

Listagem 7 – Criação de Sequências para Treinamento e Validação:

```
X_train, y_train = create_sequences(train_data, seq_length)
X_validation, y_validation = create_sequences(validation_data, seq_length)
```

Fonte: Autoria própria (2024).

Listagem 8 – Construção do Modelo LSTM

```
 model = tf.keras.models.Sequential([
    tf.keras.layers.LSTM(64, input_shape=(seq_length, df.shape[1]),
    return_sequences=True),
    tf.keras.layers.LSTM(32, return_sequences=False),
    # return_sequences=False na última camada LSTM
    tf.keras.layers.Dense(df.shape[1])
])
```

Fonte: Autoria própria (2024).

Listagem 9 – Compilação do Modelo

```
model.compile(optimizer='adam', loss='mse')
```

Fonte: Autoria própria (2024).

Listagem 10 – Treinamento do Modelo

```
history = model.fit(X_train, y_train,
epochs=50, batch_size=32,
validation_data=(X_validation, y_validation),
verbose=2)
```

Fonte: Autoria própria (2024).

Listagem 11 – Avaliação do Modelo

```
train_loss = model.evaluate(X_train, y_train, verbose=0)
validation_loss = model.evaluate(X_validation, y_validation, verbose=0)
print(f'Train Loss: {train_loss}, Validation Loss: {validation_loss}')
```

Fonte: Autoria própria (2024).

Listagem 12 – Previsões para o Período de Validação

```
 validação
validation_start_date = pd.Timestamp('2023-10-21')
validation_end_date = pd.Timestamp('2023-10-31')

validation_input = df_normalized.
loc[validation_start_date - pd.Timedelta(days=seq_length)
:validation_end_date]
X_validation_input, _ = create_sequences(validation_input, seq_length)

validation_preds = model.predict(X_validation_input)
```

Fonte: Autoria própria (2024).

Listagem 13 – Desnormalização das Previsões

```
# Desnormalizar as previsões
validation_preds_df = pd.DataFrame(scaler.
inverse_transform(validation_preds), columns=df.columns,
index=validation_input.index[seq_length:])
```

Fonte: Autoria própria (2024).

Listagem 14 – Plotagem das Previsões Gerais

```
# Plotar o gráfico geral de previsões
plt.figure(figsize=(12, 6))

for column in df.columns:
    plt.plot(validation_input.index[seq_length:], validation_input[column]
    [seq_length:], label=f'{column} Real')

for column in df.columns:
    plt.plot(validation_preds_df.index, validation_preds_df[column],
    label=f'{column} Previsto')

plt.xlabel('Data')
plt.ylabel('Valores')
plt.title('Valores Reais vs. Valores Previstos (Período de Validação)')
plt.legend()
plt.grid(True)
plt.show()
```

Fonte: Autoria própria (2024).

Listagem 15 – Plotagem de Gráficos Individuais

```
# Plotar gráficos individuais para cada gás
for column in df.columns:
    plt.figure(figsize=(12, 6))
    plt.plot(validation_input.index[seq_length:], validation_input[column]
    [seq_length:], label='Real', color='blue')
    plt.plot(validation_preds_df.index, validation_preds_df[column],
    label='Previsto', color='red')
    plt.xlabel('Data')
    plt.ylabel('Valores')
    plt.title(f'Valores Reais vs.
    Valores Previstos para {column}
    (Período de Validação)')
    plt.legend()
    plt.grid(True)
    plt.show()
```

Fonte: Autoria própria (2024).