Impacto da Imputação de Dados na Predição da Qualidade do Ar: Um Estudo de Caso em Congonhas-MG

João A. S. Silva[1,2], **Felipe D. Cunha**[2]

[1] Instituto de Computação – (UNICAMP)
Campinas – São Paulo – Brasil

[2]Instituto de Ciências Exatas e Informática – (PUC Minas)
Belo Horizonte – Minas Gerais – Brasil

j270804@dac.unicamp.br, felipe@pucminas.br

Abstract. *The number of people affected by diseases related to poor air quality has increased significantly over the years, totaling approximately* 6.7 *million deaths annually worldwide. However, there is still a lack of specific applications focused on predicting air quality to warn the population about imminent risks. Given this scenario, the literature presents various machine-learning techniques that can be used to forecast air quality. Nevertheless, the databases must be complete without missing values for these techniques to be effective. This work investigates the impact of imputing missing data on air quality prediction in the city of Congonhas-MG. The results indicate that, although there are simple imputation methods and algorithms capable of handling incomplete data, applying appropriate techniques to fill these gaps significantly improves the accuracy of air quality predictions. This enables more efficient warnings to the population about the risks associated with exposure to air pollutants.*

Resumo. *O número de pessoas afetadas por doenças relacionadas à baixa qualidade do ar tem aumentado significativamente ao longo dos anos, totalizando aproximadamente* $6,7$ *milhões de mortes anuais em todo o mundo. No entanto, ainda há uma carência de aplicações específicas voltadas para a previsão da qualidade do ar com o objetivo de alertar a população sobre riscos iminentes. Diante desse cenário, a literatura apresenta diversas técnicas de aprendizado de máquina que podem ser utilizadas para prever a qualidade do ar. Contudo, para que essas técnicas sejam eficazes, é fundamental que as bases de dados estejam completas, sem a presença de valores ausentes. Este trabalho investiga o impacto da imputação de dados ausentes na predição da qualidade do ar na cidade de Congonhas-MG. Os resultados obtidos indicam que, embora existam métodos simples de imputação e algoritmos capazes de lidar com dados incompletos, a aplicação de técnicas adequadas para preencher essas lacunas tende a melhorar significativamente a precisão das predições da qualidade do ar. Isso possibilita alertar a população de forma mais eficiente sobre os riscos associados à exposição a poluentes atmosféricos.*

1. Introdução

Nos últimos anos, o assunto de qualidade do ar tem se tornado importante pauta para discussão, tendo em vista o grande impacto negativo na saúde das pessoas e no meio

ambiente, tornando-se perceptível e se agravando cada vez mais, sendo responsável pela causa de problemas no sistema nervoso central, problemas respiratórios, cardiovasculares, renais e reprodutórios. Além de ser um dos principais agravantes dos problemas de saúde, a baixa qualidade do ar nos ambientes internos e externos combinados é responsável pela morte de 6.7 milhões de pessoas ao redor do mundo [World Health Organization 2022]. A partir desses dados, a compreensão sobre as fontes de emissão de gases poluentes, as suas propriedades particulares e possíveis danos à saúde podem ser um fator chave para auxiliar as autoridades competentes a realizarem fiscalizações rigorosas para melhor controle do meio. Um exemplo de tentativa de fiscalização é a implementação da Lei 14.850/24[1], que prevê a criação de padrões nacionais da qualidade do ar no Brasil.

Atualmente, existem estudos na área de Cidades Inteligentes voltados para a medição da qualidade do ar em centros urbanos, mensurando e analisando poluentes emitidos por veículos e indústrias, assim como é feito em [Campos et al. 2021]. Nele, os autores propõem a utilização de modelos de sensoriamento virtual para prever a qualidade do ar com base em dados advindos de sensores físicos espalhados em várias cidades do mundo, incluindo São Paulo, cidade brasileira conhecida pelo alto número de veículos em circulação e, consequentemente, por uma qualidade do ar mais baixa. Além do trânsito ser um grande agressor à qualidade do ar na cidade de São Paulo, outras cidades do Brasil sofrem com a forte poluição causada pelas atividades de extração mineral, como é o caso da cidade de Congonhas, no estado de Minas Gerais.

A cidade de Congonhas é conhecida por vários fatores, pelo seu centro histórico que atrai milhares de turistas anualmente e principalmente pelas atividades de grandes mineradoras em seu território, sendo responsável por 85% da arrecadação do município, fazendo com que Congonhas seja uma das protagonistas na extração de minério no Brasil. Com a alta quantidade de minério extraído, existem consequências ambientais que se agravam com o aumento das atividades extrativistas, sendo elas, a presença de barragens de rejeito de minério, que totalizam 24 na cidade de Congonhas, sendo 17 em atividade e 7 desativadas. Além do risco trazido pela presença das barragens, Congonhas apresenta um grave problema em relação à qualidade do ar da cidade, assim como é apresentado pelos autores em [Andrade et al. 2016], devido à constante extração de minério em minas abertas no município, que constantemente sofre com nuvens de poeira no ar, assim como é mostrado em [Jornal Estado de Minas 2021]. Estas nuvens são um indicativo visual da baixa qualidade do ar, devido a alta quantidade de poluentes no ar, em especial as Partículas Inaláveis de diâmetro inferior a 10 micrômetros (μm) PM10, as Partículas Inaláveis de diâmetro inferior a 2,5 micrômetros (μm) PM2.5 e Partículas Totais em Suspensão (PTS), o que causa preocupação aos moradores e à sua saúde respiratória. Além do indicativo visual da poluição do ar, Congonhas apresenta indicadores da qualidade do ar em índices elevados, apontando risco à saúde das pessoas.

O presente trabalho busca analisar e predizer a qualidade do ar na cidade de Congonhas para alertar a população sobre possíveis riscos à saúde, fazendo uso de técnicas e algoritmos de Inteligência Artificial para tratar e analisar dados provenientes de estações de monitoramento da qualidade do ar instaladas na cidade de Congonhas-MG, propostas em [Luiz C. D. Santolim 2017], que possibilitam o monitoramento de hora em hora de parâmetros relevantes para classificar o ar do município. Além disso, visando garantir a

[1]https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/L14850.htm

qualidade dos dados, o objetivo central do presente trabalho é realizar de forma efetiva o tratamento de dados ausentes e apresentar o impacto de tal tratamento na predição final a ser realizada, tendo em vista a alta quantidade de valores ausentes nas bases de dados em questão, devido a diversos fatores, o principal deles, o mau funcionamento dos sensores responsáveis por medir a qualidade do ar neste município.

O trabalho está organizado da seguinte maneira, na Seção 2 são abordados os trabalhos que possuem maior relevância na motivação do trabalho e na metodologia escolhida. Na Seção 3 é discutida a metodologia com a descrição sobre a coleta de dados, as bases de dados utilizadas, etapas iniciais do pré-processamento dos dados escolhidos para análise, incluindo a imputação de dados ausentes e a descrição do processo para a predição da qualidade do ar. Na Seção 3.4 são abordadas as técnicas para imputação de dados ausentes utilizadas no presente trabalho e a comparação de métricas obtidas após a execução das mesmas. Na Seção 4 é feita a comparação de métricas obtidas após a imputação de dados e os resultados obtidos na predição da qualidade do ar, por fim, na Seção 5 é apresentada a conclusão do trabalho e a proposição de trabalhos futuros.

2. Trabalhos Relacionados

Em [Braga et al. 2007], os autores investigam os efeitos da exposição de crianças e idosos ao material particulado gerado pela mineração na cidade de Itabira, Minas Gerais. O estudo analisa os atendimentos de pronto-socorro relacionados a doenças respiratórias entre os anos de 2003 e 2004. Os resultados indicam que a poluição do ar em Itabira está associada ao aumento nos atendimentos de pronto-socorro por doenças respiratórias em crianças e adolescentes, bem como por doenças cardiovasculares em adultos. Além disso, o trabalho destaca que os efeitos respiratórios tendem a ser mais prolongados em comparação com os efeitos cardiovasculares.

Já em [Luiz C. D. Santolim 2017], os autores apresentam estudos realizados na cidade de Congonhas-MG, advindos de parcerias entre a Secretaria do Meio Ambiente da cidade, o Ministério Público de Minas Gerais (MPMG), a Fundação Estadual do Meio Ambiente (FEAM) e mineradoras da região, a fim de compreender a qualidade do ar na cidade e as condições meteorológicas na mesma. Foram utilizadas técnicas de modelagem adotadas no mundo todo, como o WRF (*Weather Research and Forecasting Model*) e CMAQ (*Community Multiscale Air Quality Modeling System*), que buscam compreender as emissões atmosféricas nas áreas de estudo e identificar os poluentes que mais agridem a qualidade do ar. Após a execução dos estudos na cidade de Congonhas-MG, os autores concluem que Material Particulado (MP), ou Particulate Matter (PM) em inglês, é o poluente que apresenta maior impacto no ar da cidade. Além disso, os autores apresentam um plano de monitoramento meteorológico e de poluentes no ar, contendo 16 estações de monitoramento, sendo 8 estações voltadas para o monitoramento de poluentes, que serão objeto de estudo do presente trabalho, e outras 8 estações para monitorar as condições climáticas da região.

Em [Anil Jadhav and Ramanathan 2019], os autores realizam a comparação entre métodos de imputação de dados ausentes em bases numéricas de diferentes contextos e comparam os resultados através da métrica *Normalized Root Mean Squared Error (NRMSE)*. Para tanto, os autores fazem a comparação entre métodos de imputação simples e imputação múltipla, destacando as principais diferenças entre os algoritmos em questão

e também os seus resultados dentro das bases de dados analisadas. Em suma, os resultados obtidos pelos autores apontam que o algoritmo *KNN Imputer* obteve as melhores métricas após análise, o que incentivou a utilização do mesmo no presente trabalho, além da métrica *NRMSE* para a comparação dos resultados entre os diferentes métodos de imputação.

Por fim, em [Doreswamy et al. 2017], os autores discutem diferentes formas para imputação de dados ausentes, em bases de dados contendo informações climáticas coletadas em 79 estações de monitoramento na Índia, disponibilizadas pelo *National Climatic Data Center (NCDC)*. A base de dados utilizada pelos autores contém informações de 2000 até 2016, apresentando quantidade significativa de dados ausentes. A imputação de dados ausentes é uma tarefa essencial no pré-processamento de dados para a execução de algoritmos de inteligência artificial, onde, a depender do valor escolhido para substituir um valor ausente, os resultados do modelo escolhido serão prejudicados, não entregando valores condizentes com a realidade.

O presente trabalho se diferencia dos demais por propor uma metodologia capaz de tratar os dados ausentes das coletas nas estações de monitoramento e realizar a predição da qualidade do ar com base nos poluentes medidos nas estações de monitoramento propostas em [Luiz C. D. Santolim 2017]. Além disso, o presente trabalho faz uso da metodologia de pré-processamento dos dados, assim como proposto em [Doreswamy et al. 2017] e em [Anil Jadhav and Ramanathan 2019], para, desta forma, permitir que os dados sejam utilizados para a predição da qualidade do ar proposta na Seção 4, que tem como objetivo principal alertar a população sobre os riscos à saúde causados pela poluição do ar apontados em [Braga et al. 2007], [Liu et al. 2022] e em [Lelieveld et al. 2023].

3. Metodologia

Pode-se observar na Figura 1 a metodologia adotada para o desenvolvimento deste trabalho. Para tanto, a etapa de pré-processamento se mostrou como uma das mais importantes para a condução dos estudos. Nas próximas Seções, serão apresentadas todas as etapas necessárias para a condução e avaliação deste estudo.

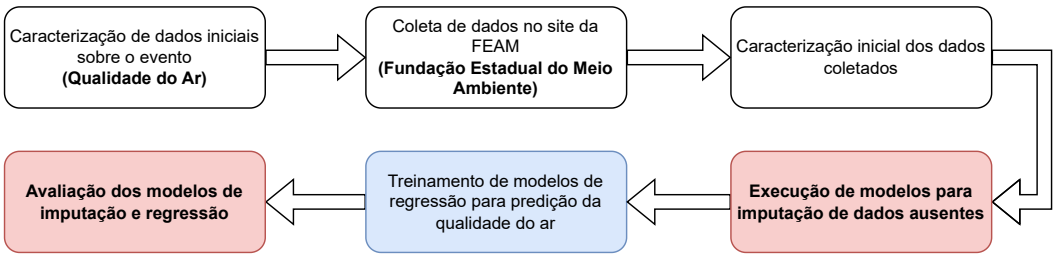

Figura 1. Fluxograma das Etapas do Trabalho.

3.1. Base de Dados

A Fundação Estadual do Meio Ambiente do estado de Minas Gerais (FEAM)[2] tem como objetivo promover a preservação, a conservação e a recuperação dos ecossistemas no estado, sendo também responsável por destinar recursos financeiros para essa finalidade e fiscalizar assuntos relacionados ao meio ambiente. Seguindo a política de Dados Abertos,

[2]https://feam.br/

instituída no Governo Federal do Brasil a partir do decreto nº 8.777 [3], o estado de Minas Gerais está adaptando os seus órgãos a se adequarem ao portal de dados abertos no estado, o que tem levado dados da FEAM a se tornarem disponíveis para todos. Tendo isso em vista, estão disponíveis no site da FEAM para livre acesso dados referentes à qualidade do ar na cidade de Congonhas-MG, coletados em 8 estações espalhadas pela cidade, com dados a partir de 2017 e finalizando em 2023.

A Tabela 1 mostra as 8 estações de monitoramento do ar em Congonhas-MG, a tabela também mostra o tipo de dado coletado em cada estação, nas quais as estações de qualidade são aquelas voltadas a monitorar a quantidade de gases poluentes suspensos no ar da cidade. Já as estações classificadas como meteorológicas apresentam apenas dados meteorológicos, como por exemplo, temperatura, velocidade do vento, umidade relativa do ar, entre outros atributos. Além disso, a Tabela 1 também mostra o número de instâncias coletadas para cada uma das estações, a porcentagem de dados ausentes e a distância média para todas as minas da cidade de Congonhas. As informações sobre as minas foram extraídas no site da Companhia Siderúrgica Nacional[4].

Estação	Instâncias	Ausentes	Distância Minas	Tipo estação	PM10	PM2.5
Basílica	55,736	8.0%	6.51 km	Qual.	X	X
Casa	61,344	17.0%	6.56 km	Met.		
Jardim	52,584	48.0%	7.65 km	Qual.	X	
Lobo	58,408	20.0%	10.29 km	Qual.	X	X
Matriz	48,287	30.0%	6.35 km	Qual.	X	X
Ferrous	52,584	36.0%	6.98 km	Met.		
Novo	57,000	13.0%	6.31 km	Qual.	X	X
Pires	61,343	26.0%	8.51 km	Qual.	X	

Tabela 1. Estações de monitoramento do ar na cidade de Congonhas.

Para exploração e entendimento dos dados coletados de cada estação, todas as bases voltadas para a medição da qualidade do ar foram utilizadas para que uma pudesse ser selecionada para conduzir os estudos. No primeiro momento, foi definido que os poluentes a serem priorizados no estudo são as Partículas Inaláveis de diâmetro inferior a 10 micrômetros (µm) PM10, as Partículas Inaláveis de diâmetro inferior a 2,5 micrômetros (µm) PM2.5 e Partículas Totais em Suspensão (PTS). Essa escolha se deu devido ao grande impacto desses poluentes à saúde humana, podendo ser responsáveis por graves doenças respiratórias, conforme apontado em [Morozesk et al. 2021]. A fim de selecionar apenas uma das estações para conduzir os estudos deste trabalho, foram realizadas análises preliminares para compreender a integridade dos dados em cada uma das bases. A Tabela 1 apresenta as métricas geradas consideradas para a tomada de decisão. Foi analisado o número de instâncias em cada base, assim como pode-se observar na Figura 2, a quantidade de dados ausentes e a distância média das estações de monitoramento para as minas em Congonhas. Após a análise de todas essas variáveis, foi definido o uso dos dados da Estação Basílica para a realização dos estudos no presente trabalho.

A Figura 3 mostra uma pequena janela de dados com os dados coletados entre

[3]https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm

[4]https://www.csn.com.br/mineracao/

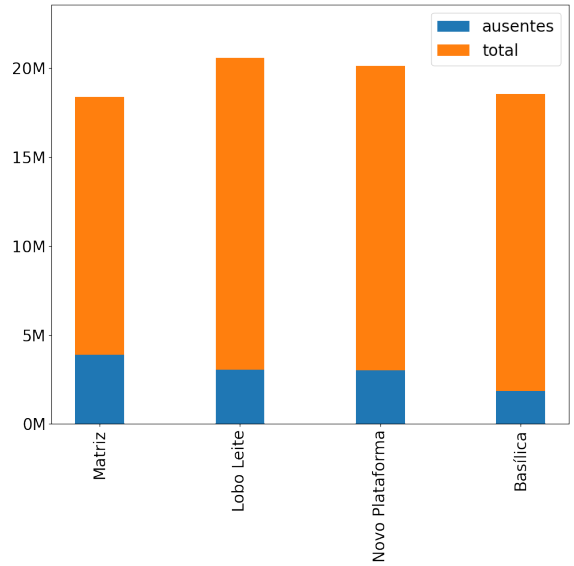

Figura 2. Quantidade de dados ausentes e o total em cada estação de monitoramento.

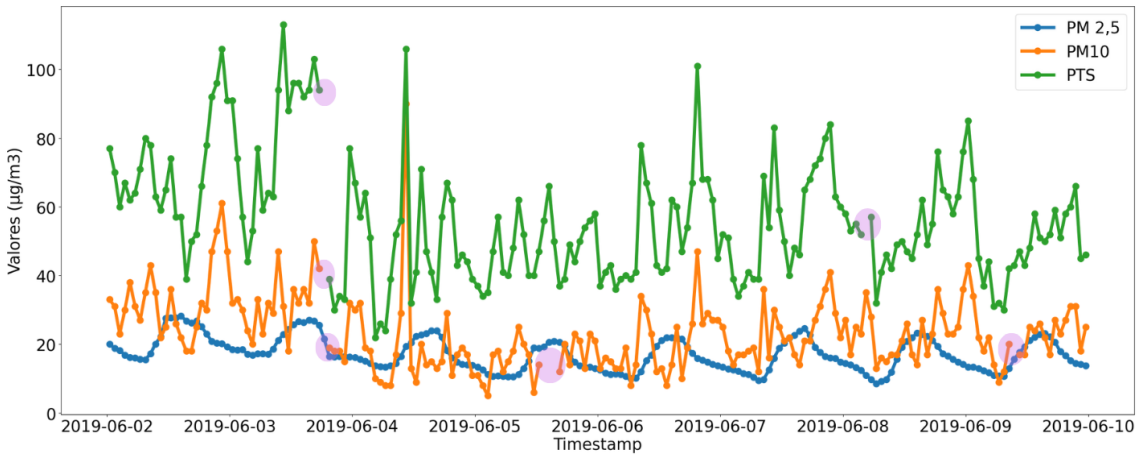

Figura 3. Janela de dados entre os dias 02/06/2019 e 09/06/2019.

os dias $02/06/2019$ e $09/06/2019$, mostrando a grande variação entre os mínimos e máximos dos valores coletados para os poluentes PM10, PM2.5 e PTS, o que dificulta bons resultados de algoritmos de predição e de imputação na base de dados em questão. Também é destacado na Figura 3 pontos onde é possível observar a ausência de dados, comprometendo a qualidade dos dados na base de dados e tornando necessária a execução de etapas de pré-processamento mais robustas de forma a reduzir o impacto dos valores ausentes através de análises para compreender o tipo de dado ausente e, em sequência, realizar a imputação desses valores, conforme é feito na Seção 3.4.

3.2. Ferramentas

Neste trabalho, foram utilizadas diversas ferramentas para a condução de cada uma das etapas do estudo. Para a coleta das bases de dados no site da FEAM[5], foi desenvolvido

[5]https://feam.br/

código utilizando a linguagem de programação *Python* para fazer o download de todos os arquivos correspondentes a cada estação de monitoramento, totalizando 54 arquivos. Após o download de todos os arquivos disponíveis, foi desenvolvido um código em *Python* para unir todos os arquivos correspondentes à mesma estação de monitoramento do ar, reduzindo os 54 arquivos para apenas 8, contendo todos os dados extraídos da FEAM. Por fim, após a etapa de unir as bases de dados, foi utilizado o *Google Colab*[6] para a execução do restante das etapas de pré-processamento necessárias, além da execução dos algoritmos de imputação de dados e de regressão para predição de valores futuros. Todos os códigos utilizados no trabalho estão disponibilizados em um repositório no *GitHub*[7].

3.3. Pré-processamento

Considerando que a base de dados da estação de monitoramento do ar Basílica contém mais informações do que aquelas necessárias para o desenvolvimento das análises deste trabalho, tornou-se necessária a remoção de atributos irrelevantes. Para tanto, filtros foram aplicados para realizar a remoção de colunas que apresentavam apenas dados ausentes e que poderiam prejudicar os resultados deste trabalho. Além disso, foi definido o tipo de dado de cada coluna, de forma a padronizar a visualização de todas as estações analisadas. A base de dados da estação Basílica contempla o *Timestamp* do momento da leitura, a quantidade de Partículas Inaláveis de 10 μm no ar (PM10), quantidade de Partículas Inaláveis de 2,5 μm no ar (PM2.5) e a quantidade de Partículas Totais em Suspensão (μm) no ar (PTS). Além disso, a base de dados contém $46,976$ instâncias, com 8% dos dados ausentes, sendo o foco do trabalho encontrar os valores mais condizentes para preencher os espaços vazios da base, imputação realizada na Seção 3.4. Após a imputação dos valores ausentes, é realizada a predição de valores futuros para os poluentes PM10, PM2.5 e PTS na Seção 4, onde o atributo *timestamp* é essencial por os dados serem distribuídos em uma série temporal.

3.4. Imputação de Dados Ausentes

A imputação de dados ausentes é uma etapa fundamental no processo de preparação dos dados e é considerada como um dos maiores desafios na análise de dados. A ausência de dados em uma base de dados se dá por diferentes fatores, podendo ser a falha de sensores, erros de coleta, perda de dados, falha humana, entre outras razões, o que pode levar a aumentos significativos no uso de recursos computacionais e distorção de resultados, assim como mostram os autores em [Alwateer et al. 2024]. A qualidade dos dados é uma característica importante, assim como dito em [Khan and Hoque 2020], onde os autores propõem o algoritmo *Single Center Imputation from Multiple Chained Equation (SICE)*, combinação entre técnicas de imputação simples, que atribuem apenas um valor para cada dado ausente, e imputação múltipla [Rubin 1996], [Rubin 1987], que atribui vários valores para cada dado ausente e realiza o processo de *pooling* para definir o valor que mais se encaixa.

Após análise inicial e observação da Figura 3, pode-se afirmar que os dados da qualidade do ar pertencem ao grupo *Missing Completely At Random (MCAR)*, conforme descrito pelos autores em [Little 1988], já que a ausência de um dado não tem relação com outros dados, assim como mostra a Figura 3, o que dificulta para os algoritmos de imputação

[6]https://colab.google.com

[7]https://github.com/joaoaugustoss/Air-Quality-Data-Imputation

obterem resultados mais precisos. A classificação dos dados ausentes presentes na base de dados limita os algoritmos capazes de fazer boas inferências para o preenchimento das lacunas existentes. A fim de comparar os resultados obtidos com cada uma das técnicas de imputação de dados, foram selecionadas janelas de dados contínuos, ou seja, sem nenhuma ocorrência de dados ausentes, e após a seleção dessa janela, no início do ano de 2022, foram imputados 20% de dados ausentes. Para diversificar os testes, foram criadas janelas de 24, 48 e 72 instâncias no início de 2022, resultando em 5, 10 e 15 dados ausentes em cada uma das janelas, respectivamente. Por fim, foram utilizadas a métricas responsáveis por indicar o erro acumulado dos modelos, *Mean Squared Error (MSE)* e *Mean Absolute Error (MAE)* que possuem como valor esperado 0 e a métrica R^2 *Score*, também conhecida como Coeficiente de Determinação, onde quanto mais o valor se aproxima de 1, mais o modelo se ajusta aos dados e caso o seu valor seja negativo, significa que os resultados do modelo são piores que preencher os valores pela média do atributo. As métricas foram calculadas com a partir do valor preenchido pela imputação e no *ground truth* gerado. Os algoritmos de imputação utilizados foram o *KNN Imputer*, método de imputação simples e o *Multivariate Imputation by Chained Equations (mice)* [van Buuren and Groothuis-Oudshoorn 2011], método de imputação múltipla.

3.4.1. KNN Imputer

O *KNN Imputer* é um algoritmo desenvolvido para a imputação de dados simples, baseado no *K-Nearest Neighbor*, que utiliza o princípio da proximidade entre as instâncias. Ao identificar os k vizinhos mais próximos de uma instância com um valor faltante, o algoritmo calcula a distância desses vizinhos com base em alguma função de distância e utiliza esse valor para preencher a lacuna. A escolha do valor de k e da métrica de distância são cruciais para a performance do método, pois influenciam diretamente a qualidade da imputação. Neste trabalho, foi definido o cálculo dos elementos com base na distância euclidiana e o número de vizinhos a serem analisados k como 48, ou seja, é imputado um elemento com base nas 48 horas conhecidas ao redor do valor faltante, a fim de trazer mais precisão ao modelo. Também foi definido o peso para o cálculo do valor ausente como a distância. Desta forma, os dados mais próximos terão mais influência no cálculo do novo valor em comparação com instâncias mais distantes.

A partir das definições apresentadas acima, o *KNN Imputer* foi executado em 3 cenários diferentes, utilizando as bases janelas selecionadas para obtenção de métricas, permitindo a observação dos resultados do modelo. Os resultados das execuções do *KNN Imputer* são exibidos na Seção 4.1.

3.4.2. *Multivariate Imputation by Chained Equations*

O algoritmo *Multivariate Imputation by Chained Equations (MICE)* é um método de imputação múltipla, capaz de estimar vários valores para um mesmo dado ausente. O *MICE* é conhecido por trabalhar bem com bases de dados complexas e também com dados referentes a qualidade do ar, como concluem os autores em [Hua et al. 2024]. A principal ideia por trás do *MICE* é modelar a relação entre a variável com dados faltantes e as outras variáveis da base de dados. Para cada variável com dados faltantes, o modelo de

regressão é ajustado, utilizando as outras variáveis como preditores. Em seguida, os valores faltantes são imputados a partir desses modelos. Esse processo é iterado k vezes, até que a convergência seja alcançada.

Assim como feito com o *KNN Imputer*, o algoritmo *MICE* também foi executado em 3 cenários distintos, utilizando as mesmas janelas de dados, a fim de gerar métricas para comparar a execução dos diferentes modelos de imputação utilizados neste trabalho. Foram utilizadas 10 iterações do *MICE* para a execução do algoritmo. A comparação dos resultados é descrita na Seção 4.1.

3.5. Predição de Valores Futuros

Com o intuito de avaliar as diferentes técnicas de imputação de dados apresentadas na Seção 3.4, nesta Seção é detalhada a execução do algoritmo de regressão *Random Forest Regressor* [Liaw and Wiener 2002]. O algoritmo é baseado na agregação de diferentes árvores de decisão que capturam padrões distintos nos dados utilizados em seu treinamento, sendo capaz de aprender sobre a influência de outros poluentes no poluente analisado e a sazonalidade dentro da série temporal. O *Random Forest Regressor* é utilizado neste trabalho para realizar a regressão da qualidade do ar, devido à sua robustez aos *outliers* e ruídos, assim como à capacidade de lidar com dados não lineares.

No desenvolvimento deste trabalho, o algoritmo *Random Forest Regressor* foi executado 2 vezes para cada poluente analisado, equivalente a uma execução por técnica de imputação utilizada para a predição do poluente PM 10, PM 2.5 e PTS, respectivamente. Para o ajuste dos parâmetros utilizados no algoritmo de regressão, foi considerado o uso do $n\_estimators = 500$, definindo a agregação de 500 árvores de decisão ao *Random Forest* e o parâmetro $max\_depth = 5$, definindo o nível máximo que cada árvore pode chegar como 5. Além disso, a fim de quantificar os resultados obtidos com a regressão, foram criados conjuntos de treino e teste com base na imputação de dados ausentes realizada na Seção 3.4, onde 80% da base de dados foi utilizada para o treinamento do regressor e os 20% restantes foram utilizados para o teste do modelo e geração das métricas analisadas. A Seção 4.2 apresenta os resultados obtidos.

4. Resultados

Nesta Seção serão avaliados os resultados obtidos com a imputação de dados e também com a predição de valores futuros com base nas bases de dados imputadas.

4.1. Imputação de Dados Ausentes

Após a execução dos algoritmos *KNN Imputer* e *MICE* para a imputação de dados ausentes, foi feito o cálculo das métricas a fim de comparar o desempenho de ambos os algoritmos. Para isso, a análise foi feita em 3 cenários distintos conforme mencionado na Seção 3.4. A Tabela 2 mostra a comparação entre as métricas obtidas com a execução de cada um dos algoritmos de imputação para a janela de dados com 24 instâncias realizando a imputação de todos os poluentes analisados. Em adição, a Tabela 3 apresenta as métricas obtidas com a execução dos algoritmos utilizando a janela de dados com 48 instâncias, mantendo a divisão entre os poluentes. Por fim, a Tabela 4 traz os resultados obtidos com a execução dos algoritmos para a janela de dados com 72 instâncias, mantendo a disposição dos resultados anteriores, exibindo os resultados para todos os poluentes.

Métricas	PM10		PM 2,5		PTS	
	KNN	MICE	KNN	MICE	KNN	MICE
MSE	**8.137**	10.916	**0.218**	1.735	**25.05**	34.916
MAE	**1.076**	1.25	**0.193**	0.495	**1.837**	2.333
R^2	-0.039	-0.394	**0.849**	-0.194	-0.156	-0.612

Tabela 2. Comparação das métricas obtidas após a execução do *KNN Imputer* e do *MICE* por meio da janela de dados com 24 instâncias.

Métricas	PM10		PM 2,5		PTS	
	KNN	MICE	KNN	MICE	KNN	MICE
MSE	**4.317**	4.437	**0.091**	0.412	27.678	**14.416**
MAE	0.834	**0.77**	**0.115**	0.241	2.189	**1.291**
R^2	**0.269**	0.248	**0.97**	0.864	-0.838	**0.042**

Tabela 3. Comparação das métricas obtidas após a execução do *KNN Imputer* e do *MICE* por meio da janela de dados com 48 instâncias.

Tendo em vista os resultados apresentados nas Tabelas 2, 3 e 4, é possível observar que, na maioria das métricas calculadas, o algoritmo *KNN Imputer* se mostrou com melhor desempenho no papel de preencher os dados ausentes em relação ao *MICE*. A diferença encontrada nos resultados dos algoritmos se dá pelo propósito de cada um deles, sendo que o *KNN Imputer* busca comparar os k vizinhos da mesma coluna e o *MICE* compara o valor ausente com todas as colunas, o que prejudica o resultado com a base de dados utilizada, pois as colunas podem apresentar valores ausentes ao mesmo tempo, fazendo com que os valores estimados se tornem imprecisos. Ambos os algoritmos apresentam bons resultados em suas imputações. Na Seção 4.2 são apresentados os resultados da predição de valores futuros utilizando as bases de dados imputadas.

4.2. Predição de Valores Futuros

As Figuras 4, 5, 6 mostram os valores estimados nas predições feitas pelo *Random Forest Regressor* para as últimas 96 horas do ano de 2022, enquanto a Tabela 5 expõe as métricas calculadas com base nos resultados destas predições. Como o *Random Forest Regressor* não funciona com dados ausentes, as predições utilizam as bases de dados imputadas pelos algoritmos *MICE* e *KNN Imputer*. Para tanto, são destacadas na Tabela 5 as melhores métricas obtidas na predição de cada um dos poluentes. Observando estes dados, é possível concluir que a imputação de dados ausentes utilizando o algoritmo *KNN Imputer* teve impacto positivo no resultado da previsão de dados futuros referentes ao volume dos poluentes encontrados no ar quando comparado com as predições obtidas a partir dos dados imputados com o *MICE*.

Ao comparar os valores reais dos poluentes e os valores obtidos com o *Random Forest Regressor* nas Figuras 4, 5, 6, é possível observar que o algoritmo de regressão conseguiu se manter na mesma tendência dos valores reais. Por mais que o *Random Forest Regressor* não tenha acompanhado os picos observados nos dados reais devido a falta de correlação entre os valores dos poluentes, ele se manteve próximo aos valores reais durante toda a janela observada, contribuindo para bons resultados das métricas *Mean Squared Error (MSE)* e *Mean Absolute Error (MAE)*, observadas na Tabela 5, que são

Métricas	PM10		PM 2,5		PTS	
	KNN	MICE	KNN	MICE	KNN	MICE
MSE	**4.285**	19.222	**0.289**	0.96	**17.246**	21.833
MAE	**0.74**	1.277	**0.196**	0.391	**1.729**	1.833
R^2	**0.272**	-2.264	**0.906**	0.688	-0.018	-0.289

Tabela 4. Comparação das métricas obtidas após a execução do *KNN Imputer* e do *MICE* por meio da janela de dados com 72 instâncias.

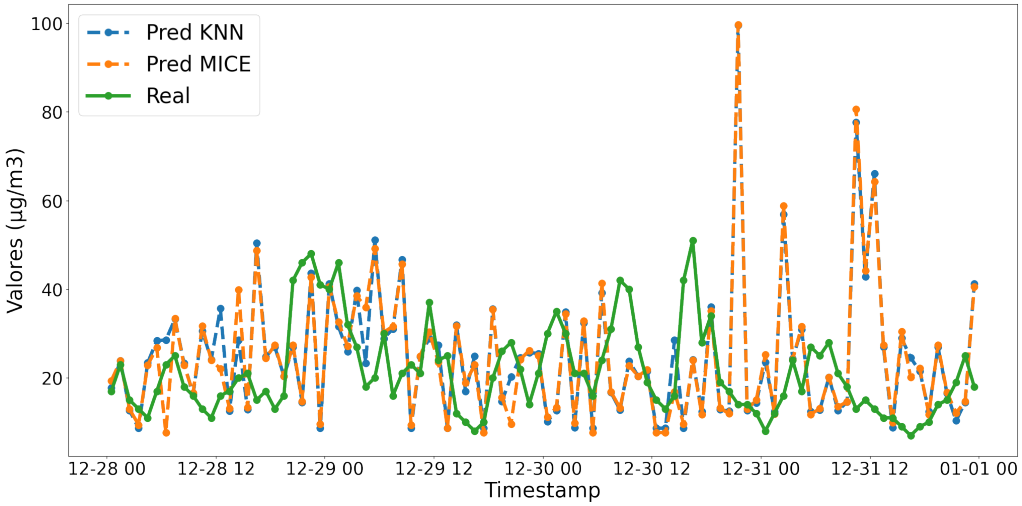

Figura 4. Comparação dos valores obtidos na execução do algoritmo *Random Forest Regressor* com as diferentes técnicas de imputação utilizadas para o PM10.

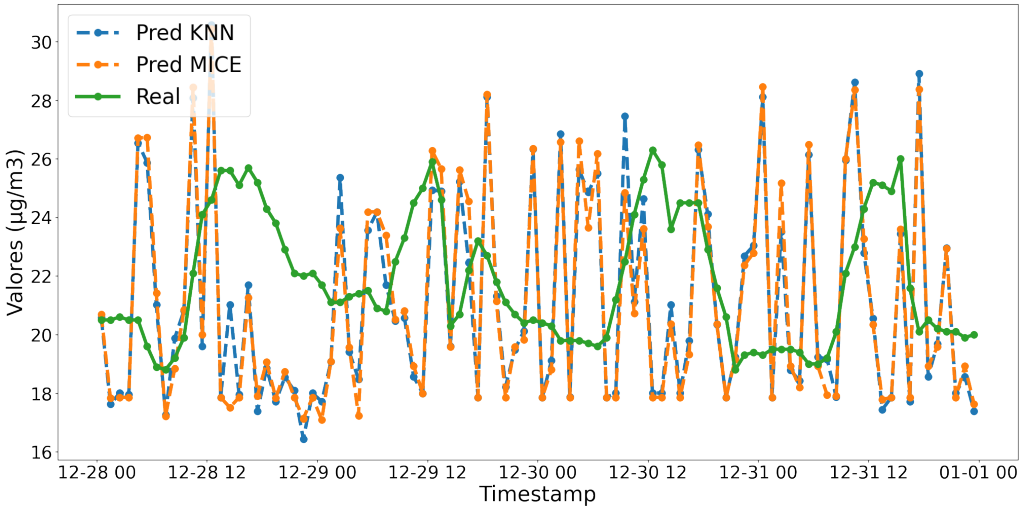

Figura 5. Comparação dos valores obtidos na execução do algoritmo *Random Forest Regressor* com as diferentes técnicas de imputação utilizadas para o PM 2,5.

calculadas a partir dos erros acumulados na fatia de dados separados para o teste do modelo de regressão. O valor alto das métricas calculadas com base no erro é explicado pelo comportamento irregular dos dados reais, observado na Figura 3.

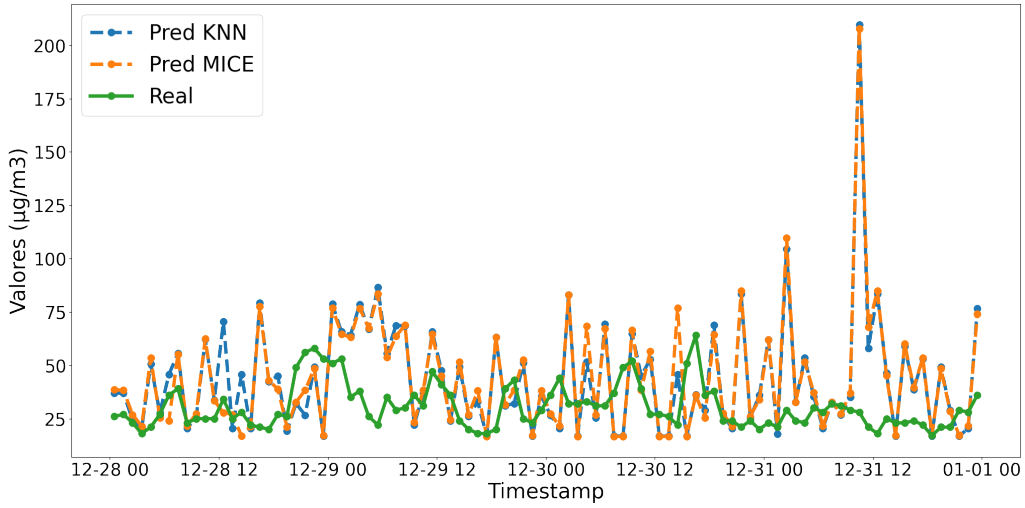

Figura 6. Comparação dos valores obtidos na execução do algoritmo *Random Forest Regressor* com as diferentes técnicas de imputação utilizadas para o PTS.

Ao observar os resultados para o poluente PTS apresentados na Figura 6, é possível notar a diferença entre os valores estimados pelo *Random Forest Regressor*, sendo que existem pontos de sobreposição entre os valores estimados pelo regressor utilizando os diferentes algoritmos de imputação utilizados. Também é possível observar pontos onde a diferença nas bases de dados influenciaram o *Random Forest Regressor* a estimar valores mais próximos dos dados reais, favorecendo às métricas expostas na Tabela 5.

Métricas	PM10		PM 2,5		PTS	
	KNN	MICE	KNN	MICE	KNN	MICE
MSE	**61.316**	74.824	**7.813**	8.763	**216.406**	257.7
MAE	**5.062**	5.628	**2.169**	2.346	**8.984**	9.917
R^2	**0.812**	0.779	**0.603**	0.570	**0.798**	0.768

Tabela 5. Comparação das métricas obtidas na execução do algoritmo *Random Forest Regressor* com as diferentes técnicas de imputação utilizadas para os poluentes PM10, PM 2,5 e PTS.

5. Conclusão e Trabalhos Futuros

Considerando os resultados obtidos com a imputação de dados ausentes, apresentados na Seção 3.4, e os resultados da predição de dados, discutidos na Seção 4, é evidente que métodos para a geração de alertas sobre a baixa qualidade do ar devem ser explorados. Esses alertas são essenciais para proteger a saúde da população de Congonhas-MG, permitindo ações preventivas antes que os níveis de poluentes atinjam patamares críticos. Dessa forma, este trabalho conduziu a análise dos dados relacionados à poluição do ar no município de Congonhas, coletados pela Federação Estadual do Meio Ambiente (FEAM-MG). Utilizando técnicas de imputação de dados, foram preenchidas as lacunas decorrentes da ausência de informações. Posteriormente, com base nos dados tratados e utilizados para o treinamento do modelo, foi realizada a predição da concentração de poluentes no ar para datas futuras.

Foram aplicadas técnicas de imputação de dados para preencher as lacunas classificadas como *Missing Completely At Random (MCAR)*. A base de dados resultante, após o processo de imputação, foi utilizada para treinar o modelo de regressão, *Random Forest Regressor*. Essa abordagem permitiu analisar e comparar as predições geradas pelo modelo com os dados reais disponíveis, avaliando tanto a eficácia do método de imputação quanto a precisão do modelo de predição. A partir dessa predição, foi possível identificar o comportamento do *Random Forest Regressor* a manter grande parte das suas estimativas dentro da mesma janela de mínimos e máximos, minimizando os erros e gerando previsões mais confiáveis, possibilitando a geração de alertas à população sobre a baixa qualidade do ar no município.

Por fim, a análise conduzida neste trabalho destaca a relevância da aplicação de técnicas de imputação de dados para garantir a qualidade dos dados, um fator essencial para alcançar resultados satisfatórios no treinamento de modelos de aprendizado de máquina. Além disso, o estudo demonstra que, com o ajuste adequado dos parâmetros e a preparação correta da base de dados, é possível realizar previsões sobre a qualidade do ar que estejam alinhadas com a realidade observada.

Para trabalhos futuros, é fundamental explorar novos algoritmos de imputação de dados, especialmente aqueles que utilizam abordagens híbridas [Khan and Hoque 2020], também a exploração de redes neurais para a predição de valores futuros, por fim, a combinação de diferentes bases de dados [Kebalepile et al. 2024] para agregação dos modelos de predição e imputação. Além disso, é recomendável realizar uma busca exaustiva pelos parâmetros mais adequados para os algoritmos de imputação e predição empregados neste estudo, visando otimizar o treinamento de modelos mais robustos e precisos. Esses avanços são essenciais para aprimorar a capacidade de alertar a população sobre possíveis riscos respiratórios decorrentes da baixa qualidade do ar.

Referências

Alwateer, M., Atlam, E.-S., Abd El-Raouf, M. M., Ghoneim, O. A., and Gad, I. (2024). Missing data imputation: A comprehensive review. *Journal of Computer and Communications*, 12(11):53–75.

Andrade, P., da Luz, J., and Campos, A. (2016). Cumulative impact assessment on air quality from multiple open pit mines. In *Clean Techn Environ Policy 18*, page 483–492.

Anil Jadhav, D. P. and Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933.

Braga, A. L. F., Pereira, L. A. A., Procópio, M., André, P. A. d., and Saldiva, P. H. d. N. (2007). Associação entre poluição atmosférica e doenças respiratórias e cardiovasculares na cidade de itabira, minas gerais, brasil. *Cadernos de Saúde Pública*, 23(suppl 4):S570–S578.

Campos, G., Cunha, F., and Villas, L. (2021). Análise de poluição atmosférica utilizando modelos de sensoriamento virtual. In *Anais do V Workshop de Computação Urbana*, pages 29–42, Porto Alegre, RS, Brasil. SBC.

Doreswamy, Gad, I., and Manjunatha, B. (2017). Performance evaluation of predictive models for missing data imputation in weather data. In *2017 International Conference*

on Advances in Computing, Communications and Informatics (ICACCI), pages 1327–1334.

Hua, V., Nguyen, T., Dao, M.-S., Nguyen, H. D., and Nguyen, B. T. (2024). The impact of data imputation on air quality prediction problem. *PLOS ONE*, 19(9):1–39.

Jornal Estado de Minas (2021). Nuvem de poeira encobre congonhas e revolta: 'muito ruim abrir os olhos'. `https://bit.ly/4iINuAY`. Acesso em: 2024-03-19.

Kebalepile, M. M., Dzikiti, L. N., and Voyi, K. (2024). Using diverse data sources to impute missing air quality data collected in a resource-limited setting. *Atmosphere*, 15(3).

Khan, S. and Hoque, A. (2020). Sice: an improved missing data imputation technique. *Journal of Big Data*, 7(37).

Lelieveld, J., Haines, A., Burnett, R., Tonne, C., Klingmüller, K., Münzel, T., and Pozzer, A. (2023). Air pollution deaths attributable to fossil fuels: observational and modelling study. *BMJ*, 383.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2:18–22. https://journal.r-project.org/articles/RN-2002-022/.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Liu, R. A., Wei, Y., Qiu, X., Kosheleva, A., and Schwartz, J. D. (2022). Short term exposure to air pollution and mortality in the us: a double negative control analysis. *Environmental Health*, 21(1):81.

Luiz C. D. Santolim, Flávio Curbani, T. J. M. (2017). Air quality assessment and design of the monitoring network of congonhas, mg, brazil. In *3rd CMAS South America - Air Pollution Conference Brazil*.

Morozesk, M., da Costa Souza, I., Fernandes, M. N., and Soares, D. C. F. (2021). Airborne particulate matter in an iron mining city: Characterization, cell uptake and cytotoxicity effects of nanoparticles from pm2.5, pm10 and pm20 on human lung cells. *Environmental Advances*, 6:100125.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

World Health Organization (2022). Ambient (outdoor) air pollution. `https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health`. Acesso em: 2024-03-19.