Interpretação de Poluentes e Variáveis Meteorológicas por Meio de Modelos Explicáveis de Aprendizado de Máquina

Guilherme S. de Farias[1], **Edler L. de Albuquerque**[2]
Marcelo A. C. Fernandes[1,3]

[1] InovAI Lab, nPITI/IMD, Universidade Federal do Rio Grande do Norte (UFRN)
Natal – RN – Brasil

[2]Instituto Federal da Bahia (IFBA) – Salvador – BA – Brasil

[3]Departamento de Engenharia de Computação e Automação (DCA)
UFRN – Natal – RN – Brasil

guilhermefariias1995@gmail.com, edler@ifba.edu.br

mfernandes@dca.ufrn.br

Abstract. *This work investigates the interpretation of atmospheric patterns based on air quality monitoring data from the city of Salvador (BA), collected between 2011 and 2016 by stations operated by CETREL S.A. The dataset includes hourly averages of meteorological variables and pollutant concentrations, totaling more than 190,000 samples after data cleaning and preprocessing. The Random Forest algorithm is used for station classification, with performance evaluated through accuracy, F1-score, and AUC-ROC. Interpretability is analyzed using Explainable Artificial Intelligence (XAI) techniques, with emphasis on SHAP, which allows identifying the contribution of variables and revealing relevant patterns for air quality analysis.*

Resumo. *Este trabalho investiga a interpretação de padrões atmosféricos a partir de dados de monitoramento da qualidade do ar da cidade de Salvador (BA), coletados entre 2011 e 2016 por estações operadas pela CETREL S.A. O conjunto inclui médias horárias de variáveis meteorológicas e concentrações de poluentes, totalizando mais de 190 mil amostras após limpeza e tratamento dos dados. Utiliza-se o algoritmo Random Forest para classificação das estações, com avaliação por acurácia, F1-score e AUC-ROC. A interpretabilidade é analisada por técnicas de Inteligência Artificial Explicável (XAI), com destaque para SHAP, permitindo identificar a contribuição das variáveis e revelar padrões relevantes para a análise da qualidade do ar.*

1. Introdução

A poluição atmosférica constitui um dos principais desafios ambientais enfrentados por centros urbanos [Tasioulis et al. 2025], em função da intensificação das atividades industriais, do crescimento da frota veicular e da complexidade dos processos físicos e químicos que governam a dinâmica dos poluentes na atmosfera. A exposição contínua a contaminantes atmosféricos está associada a impactos diretos na saúde pública, no meio ambiente e na qualidade de vida da população [Chakraborty et al. 2024], o que

torna o monitoramento e a análise da qualidade do ar elementos centrais para o planejamento urbano e a formulação de políticas públicas. Nesse contexto, compreender como variáveis meteorológicas interagem com diferentes poluentes é fundamental para interpretar padrões espaciais e temporais da poluição e apoiar processos de tomada de decisão baseados em evidências [Costa et al. 2022].

O uso de técnicas de Aprendizado de Máquina (Machine Learning – ML) tem se consolidado como uma abordagem relevante para a análise de fenômenos atmosféricos, em função da crescente disponibilidade de dados ambientais e da complexidade inerente aos processos físicos e químicos que governam a qualidade do ar. Métodos de ML têm sido empregados tanto em tarefas preditivas quanto exploratórias, permitindo a identificação de padrões não lineares, a análise de relações multivariadas entre poluentes e variáveis meteorológicas e a caracterização espacial e temporal de sistemas atmosféricos complexos. Essas abordagens oferecem vantagens em relação a métodos estatísticos tradicionais, sobretudo na capacidade de lidar com grandes volumes de dados e com interações complexas entre variáveis, tornando-se ferramentas promissoras para o monitoramento ambiental, a avaliação da qualidade do ar e o apoio à formulação de políticas públicas baseadas em dados [Rybarczyk and Zalakeviciute 2018, Méndez et al. 2023].

Trabalhos anteriores com este mesmo conjunto de dados aplicaram mapas auto-organizáveis (SOM) para analisar poluentes e variáveis meteorológicas em Salvador [Costa et al. 2022], revelando correlações e padrões de concentração. Estudo posterior ampliou a abordagem para caracterizar assinaturas atmosféricas das estações, evidenciando heterogeneidade espacial influenciada por fatores meteorológicos e emissões locais [Costa et al. 2024]. Embora robustos, esses métodos não supervisionados não aliam desempenho preditivo à interpretabilidade explícita.

Paralelamente, estudos recentes em predição da qualidade do ar têm focado no aumento da acurácia. Exemplos incluem modelos otimizados para índice de qualidade do ar na Índia [Natarajan et al. 2024] e métodos de combinação de modelos (empilhamento) para predição de poluentes [Emeç and Yurtsever 2025]. Em conjunto, esses trabalhos evidenciam o potencial do aprendizado de máquina para previsão, mas mantêm foco no desempenho preditivo, tratando os modelos como caixas-pretas [Tasioulis et al. 2025, Chakraborty et al. 2024].

Diante desse contexto, este trabalho investiga a interpretação das relações entre poluentes e variáveis meteorológicas por meio de modelos explicáveis de aprendizado de máquina, aplicados a dados reais de Salvador (2011-2016) de oito estações [Costa et al. 2022]. Utiliza-se o algoritmo Floresta Aleatória (Random Forest) para classificação das estações e o método SHAP (Inteligência Artificial Explicável) para analisar a influência das variáveis. Os resultados mostram acurácia de 0,87, acurácia balanceada de 0,86 e F1-macro de 0,86. As análises de explicabilidade revelam predominância de variáveis meteorológicas (63,5%), especialmente o desvio-padrão da direção do vento, a velocidade do vento e a precipitação, além de heterogeneidade espacial entre as estações.

2. Materiais e Métodos

2.1. Dados

A base de dados utilizada neste estudo foi obtida a partir da rede de monitoramento da qualidade do ar implantada no estado da Bahia, operada pela CETREL S.A., contem-

plando oito estações localizadas na cidade de Salvador: Av. ACM–Detran (ACM), Av. Barros Reis (BR), Paralela–CAB (CAB), Campo Grande (CG), Dique do Tororó (DT), Itaigara (IT), Pirajá (PI) e Rio Vermelho (RV) [Souza de Farias et al. 2026]. O objetivo do sistema de monitoramento é acompanhar a variabilidade espaço-temporal dos poluentes atmosféricos e sua relação com fatores meteorológicos em um grande centro urbano. As medições foram realizadas continuamente ao longo do período de 2011 a 2016, resultando em uma base extensa e representativa das condições ambientais da cidade, conforme também explorado em estudos anteriores com o mesmo conjunto de dados [Costa et al. 2022, Costa et al. 2024].

O conjunto de dados é composto por médias horárias de variáveis meteorológicas e concentrações de poluentes atmosféricos, totalizando inicialmente 420.864 amostras. As variáveis meteorológicas incluem velocidade do vento (WIND_SPEED), temperatura do ar (TEMP), umidade relativa (HUM), desvio-padrão da direção do vento (STWD), precipitação pluviométrica (RAIN) e direção do vento (WIND_DIR). Os poluentes monitorados são dióxido de enxofre (SO_2), monóxido de carbono (CO), ozônio (O_3), óxidos de nitrogênio (NOx), material particulado com diâmetro aerodinâmico inferior a 10 µm (MP10), além dos gases NO e NO_2. Essas variáveis refletem tanto processos de emissão primária, associados principalmente ao tráfego veicular e atividades industriais, quanto processos fotoquímicos e de dispersão atmosférica, fortemente influenciados pelas condições meteorológicas.

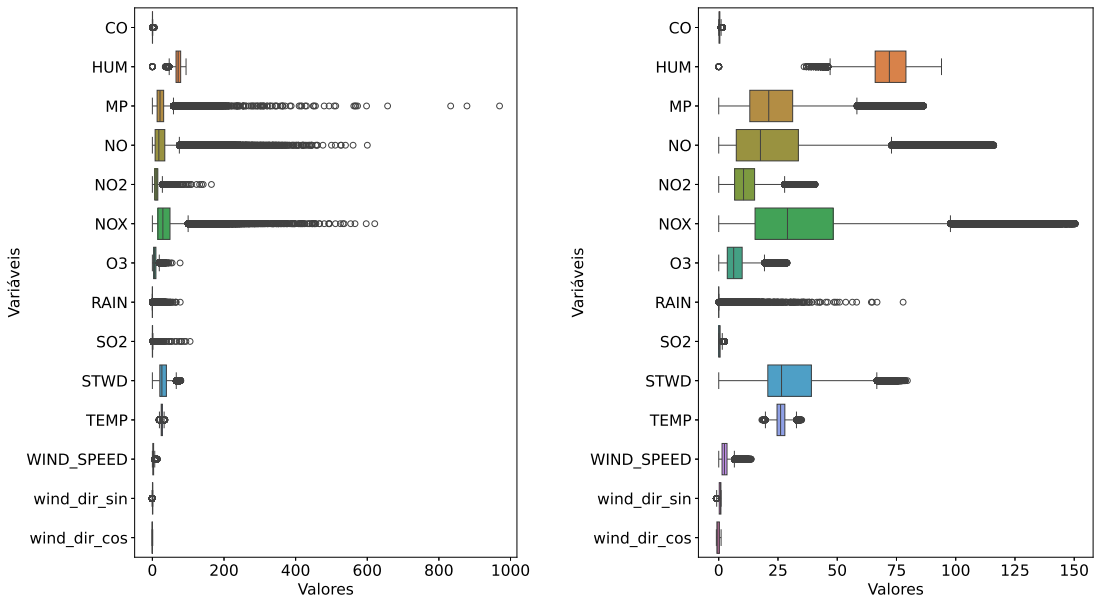
2.2. Pré-processamento dos Dados

O pré-processamento dos dados foi realizado com o objetivo de assegurar consistência e adequação do conjunto de dados às etapas posteriores de modelagem. Inicialmente, foram removidas todas as amostras que continham valores ausentes ou inconsistências associadas a erros de leitura dos sensores. Em seguida, foi conduzida uma análise exploratória das distribuições das variáveis por meio de *boxplots*, permitindo a identificação de padrões de dispersão e valores extremos, especialmente nas séries dos poluentes atmosféricos. A variável direção do vento (*WIND_DIR*) foi tratada de forma específica devido à sua natureza angular, com valores no intervalo de 0° a 360°. Essa representação pode introduzir descontinuidades artificiais em métricas de distância e afetar o desempenho de modelos de aprendizado de máquina. Para evitar esse problema, a direção do vento foi representada por meio de suas componentes trigonométricas, utilizando o seno e o cosseno do ângulo. Dessa forma, foram criadas as variáveis *WIND_DIR_sin* e *WIND_DIR_cos*, e a variável original foi removida do conjunto de dados.

A Figura 1a apresenta os *boxplots* das variáveis meteorológicas e dos poluentes atmosféricos antes da aplicação dos critérios de filtragem, permitindo visualizar a distribuição dos dados, a assimetria e a presença de valores extremos. Observa-se que as variáveis associadas aos poluentes exibem maior dispersão e um número significativo de outliers, especialmente em NO, NOx, CO e MP_{10}, refletindo episódios de alta concentração e possíveis leituras espúrias. Em contraste, as variáveis meteorológicas apresentam distribuições mais concentradas, com exceção de RAIN e STWD, cujos valores extremos estão associados à variabilidade natural dos fenômenos atmosféricos. Essa análise visual fundamentou a adoção de estratégias diferenciadas para o tratamento de outliers, restringindo a filtragem às variáveis de poluentes e preservando os extremos das variáveis meteorológicas. A identificação e remoção de outliers foi aplicada apenas

às variáveis de poluentes (CO, NO, NO_2, NO_x, O_3, MP_{10} e SO_2), utilizando o método baseado no intervalo interquartil (IQR). Para cada variável, foram calculados os quartis Q_1 e Q_3, bem como a amplitude interquartil ($IQR = Q_3 - Q_1$). Foram removidas as amostras cujos valores se encontravam fora do intervalo $[Q_1 - 1{,}5 \times IQR,\ Q_3 + 3{,}0 \times IQR]$. As variáveis meteorológicas não passaram por filtragem de outliers, uma vez que valores extremos nessas séries correspondem podem corresponder a fenômenos atmosféricos, como descrito em [Wallace and Hobbs 2006, Stull 2012, Clifton et al. 2020, Monteiro and Zanella 2023]. Após a conclusão do pré-processamento, obteve-se um conjunto final composto por 193.569 amostras, utilizado nas etapas subsequentes de treinamento e análise dos modelos.

A Figura 1b apresenta os boxplots das variáveis meteorológicas e dos poluentes atmosféricos após a aplicação dos critérios de remoção de outliers. Observa-se uma redução significativa da dispersão nas séries de poluentes, com a eliminação de valores extremos que poderiam influenciar de forma indevida o treinamento dos modelos de aprendizado de máquina. As distribuições resultantes mostram intervalos interquartis mais bem definidos e maior concentração das amostras em faixas representativas do comportamento típico das variáveis. Em contrapartida, as variáveis meteorológicas mantêm sua variabilidade característica, preservando informações associadas a fenômenos atmosféricos reais. Essa etapa assegura um conjunto de dados mais estável e adequado para a modelagem subsequente, sem comprometer a representatividade física do sistema analisado.



(a) Antes da remoção de outliers. **(b) Após a remoção de outliers.**

Figura 1. Boxplots das variáveis meteorológicas e dos poluentes atmosféricos.

Após a conclusão do pré-processamento, obteve-se um conjunto final composto por 193.569 amostras, utilizado nas etapas subsequentes de treinamento e análise dos modelos. A versão processada e utilizada neste estudo está disponível publicamente no Mendeley Data [Souza de Farias et al. 2026].

2.3. Treinamento e Validação do Modelo Random Forest

Após o pré-processamento, o conjunto de dados foi utilizado para o treinamento de um modelo supervisionado baseado no algoritmo Random Forest (RF). Esse algoritmo foi escolhido por sua capacidade de lidar com relações não lineares, por ser pouco sensível à multicolinearidade entre variáveis de entrada e por permitir a extração direta de informações sobre a contribuição das variáveis no processo de decisão [Parmar et al. 2018, Shaik and Srinivasan 2018]. O modelo foi empregado para a tarefa de classificação das estações de monitoramento, utilizando como atributos as variáveis meteorológicas e as concentrações de poluentes. A formulação do problema como uma tarefa de classificação tem como objetivo criar um cenário no qual técnicas de XAI possam identificar quais variáveis são mais relevantes para caracterizar cada estação, permitindo analisar diferenças espaciais nos padrões atmosféricos e de poluição a partir das explicações geradas pelo modelo.

A base de dados apresenta distribuição relativamente equilibrada entre as oito estações de monitoramento consideradas neste estudo. A estação Dique do Tororó (DT) concentra aproximadamente 19,2% das amostras, seguida por Paralela–CAB (16,7%) e Campo Grande (13,1%). As demais estações apresentam proporções próximas, com Pirajá (11,5%), Av. ACM–Detran (11,0%), Rio Vermelho (10,2%), Av. Barros Reis (10,0%) e Itaigara (8,4%). Embora exista variação moderada no número de amostras por estação, caracterizando um leve desbalanceamento entre as classes, a razão entre a classe mais frequente e a menos frequente permanece em um intervalo que não compromete o desempenho de algoritmos baseados em árvores, como o RF, reconhecidamente robustos a esse tipo de desbalanceamento [Parmar et al. 2018, Shaik and Srinivasan 2018]. O modelo foi treinado utilizando validação cruzada do tipo k-fold para avaliar a capacidade de generalização e reduzir a dependência de uma única partição dos dados. A seleção dos hiperparâmetros foi realizada por busca em grade (*GridSearchCV*), considerando parâmetros como número de árvores, profundidade máxima e número mínimo de amostras por nó, sendo o modelo final escolhido com base no desempenho médio obtido nas dobras de validação.

A avaliação do modelo foi realizada por meio de métricas de classificação amplamente utilizadas na literatura, incluindo acurácia, precisão, revocação (*recall*), F1-score e área sob a curva ROC (AUC-ROC). Adicionalmente, foram analisadas as matrizes de confusão para verificar o comportamento do classificador em relação às diferentes classes. Essas métricas forneceram subsídios para a comparação do desempenho preditivo do modelo e serviram como base para a etapa posterior de interpretação por meio de técnicas de XAI. A Tabela 1 apresenta um resumo da configuração experimental adotada para o treinamento e a validação do modelo RF. São descritos o conjunto de variáveis de entrada utilizadas, a divisão dos dados em conjuntos de treino e teste, bem como a estratégia de validação cruzada e de busca de hiperparâmetros empregada. A tabela também explicita as estações de monitoramento consideradas como classes no problema de classificação. Essas informações definem o cenário experimental no qual o modelo foi treinado e avaliado, assegurando reprodutibilidade e clareza quanto às decisões metodológicas adotadas.

2.4. Aplicação de Inteligência Artifical Explicável - XAI

Após o treinamento e validação do modelo RF, foram empregadas técnicas de Inteligência Artifical Explicável (XAI) para analisar o processo de decisão do classifica-

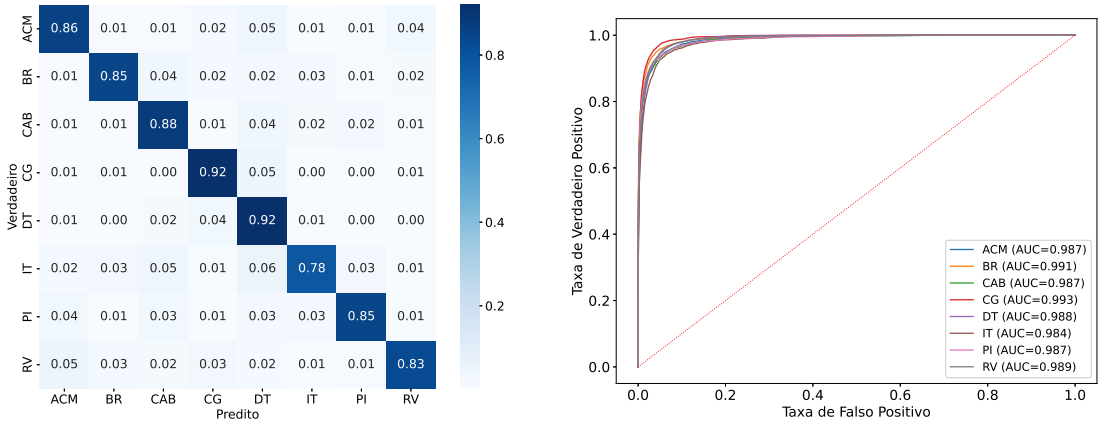
Tabela 1. Dados e estratégia de treinamento e validação do modelo RF.

Categoria	Descrição	Valor
Features utilizadas	Lista de preditores	CO, HUM, MP, NO, NO_2, NO_x, O_3, RAIN, SO_2, STWD, TEMP, WSPD, WD_{sin}, WD_{cos}
Tamanho do treino	Dados após divisão 80%	$154{,}855 \times 14$
Tamanho do teste	Dados após divisão 20%	$38{,}714 \times 14$
Validação cruzada	Estratégia de validação	5-folds (GridSearchCV)
Busca de hiperparâmetros	Total de combinações avaliadas	24 candidatos $\times$ 5 folds
Classes avaliadas	Estações de monitoramento	ACM, BR, CAB, CG, DT, IT, PI, RV

dor. Para isso, utilizou-se o método SHapley Additive exPlanations (SHAP), que quantifica a contribuição individual de cada variável de entrada para as predições do modelo [Arunika et al. 2024, Das and Rad 2020, Das et al. 2020]. O SHAP foi aplicado para obter explicações globais e locais, permitindo avaliar a importância relativa das variáveis meteorológicas e dos poluentes atmosféricos na classificação das estações de monitoramento. Baseado na teoria dos valores de Shapley, o método fornece uma decomposição consistente das predições em termos das contribuições das variáveis de entrada. Os valores de SHAP foram utilizados exclusivamente para interpretar o modelo treinado, sem interferir nas etapas de treinamento ou validação.

3. Resultados e Discussão

A Figura 2a apresenta a matriz de confusão normalizada do modelo RF, na qual os valores são expressos em termos de *recall* por classe, permitindo avaliar a taxa de acerto do classificador para cada estação de monitoramento. A Figura 2b mostra as curvas ROC no esquema *one-versus-rest* para cada classe, bem como a curva média, fornecendo uma visão global da capacidade discriminativa do modelo ao longo de diferentes limiares de decisão. Em conjunto, essas representações permitem analisar tanto o desempenho por classe quanto o comportamento geral do classificador em termos de separação entre as estações.



(a) Matriz de confusão. **(b) Curvas ROC.**

Figura 2. Avaliação do desempenho do modelo Random Forest.

O resultado apresentado através da Figura 2a indica que o modelo apresenta elevado desempenho na classificação das estações de monitoramento, com valores de *recall* elevados ao longo da diagonal principal da matriz de confusão e baixos índices

de confusão entre classes distintas. As curvas ROC (ver Figura 2b) exibem áreas sob a curva (AUC) próximas de 1 para todas as classes, evidenciando alta capacidade de discriminação do modelo independentemente do limiar adotado. Esses resultados confirmam a adequação do RF para capturar diferenças nos padrões meteorológicos e de poluição associados às estações, mesmo na presença de variabilidade espacial e leve desbalanceamento entre classes, fornecendo uma base confiável para a etapa posterior de interpretação por meio de técnicas de XAI.

A Tabela 2 apresenta um resumo quantitativo do desempenho do modelo RF após as etapas de treinamento, validação e avaliação. Os resultados indicam desempenho consistente, com acurácia de 0,869 no conjunto de teste e *balanced accuracy* de 0,860, evidenciando que o modelo mantém taxas de acerto equilibradas entre as classes, mesmo diante de um leve desbalanceamento na distribuição das estações. O valor de F1-macro igual a 0,864 reforça a capacidade do classificador em combinar precisão e *recall* de forma homogênea entre as classes. Além disso, a proximidade entre a acurácia média obtida na validação cruzada e o desempenho no conjunto de teste indica boa capacidade de generalização do modelo, corroborando os resultados observados nas análises baseadas na matriz de confusão e nas curvas ROC.

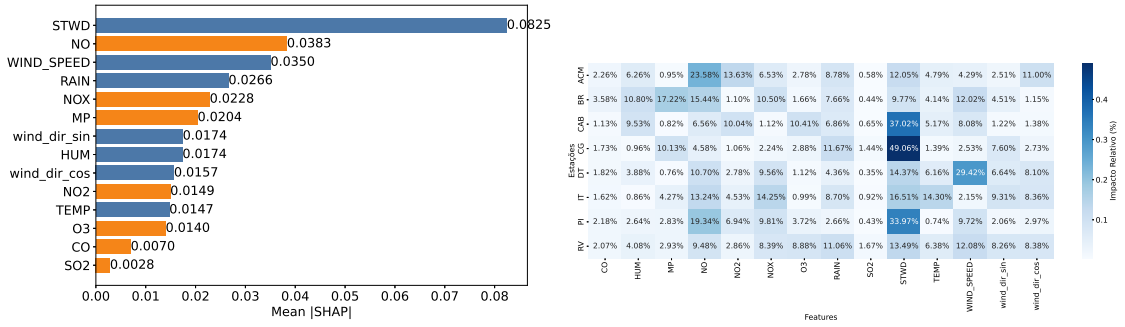
Tabela 2. Resultados de avaliação do modelo RF no conjunto de teste.

Categoria	Descrição	Valor
Busca de hiperparâmetros	Total de combinações avaliadas	24 candidatos × 5 folds
Melhor score (CV)	Acurácia média na validação cruzada	0.86176
Melhores hiperparâmetros	Configuração selecionada	`n_estimators = 500, max_depth = None` `min_samples_split = 2, min_samples_leaf = 1` `class_weight = None`
Acurácia (teste)	Acurácia global no conjunto de teste	0.86912
Balanced Accuracy	Média do *recall* ponderado	0.85954
F1-macro	Média macro do F1-score	0.86408

A Figura 3a apresenta a importância global das variáveis de entrada estimada a partir da média do valor absoluto dos SHAP values do modelo RF. A ordenação das variáveis permite identificar os atributos com maior influência nas predições do modelo, enquanto a agregação das contribuições em dois grupos — variáveis meteorológicas e poluentes atmosféricos — evidencia a participação relativa de cada conjunto no processo decisório. Observa-se que as variáveis meteorológicas apresentam maior contribuição global, representando aproximadamente 63,5% da importância total, enquanto os poluentes atmosféricos respondem por cerca de 36,5%. Entre as variáveis individuais, destaca-se o desvio-padrão da direção do vento (STWD), seguido por variáveis associadas à dinâmica do escoamento atmosférico e à precipitação, além de poluentes relacionados aos óxidos de nitrogênio. Esses resultados indicam que os processos de dispersão atmosférica desempenham papel central na diferenciação entre as estações de monitoramento, enquanto as concentrações de poluentes atuam de forma complementar, reforçando a coerência física do modelo e o potencial das técnicas de XAI para revelar relações relevantes entre meteorologia e qualidade do ar.

Já a Figura 3b apresenta a importância relativa das variáveis de entrada para cada estação de monitoramento, estimada a partir dos valores de SHAP normalizados. O mapa de calor permite comparar a contribuição das variáveis meteorológicas e dos poluentes atmosféricos no processo de decisão do modelo em diferentes estações. Observa-se que, embora algumas variáveis apresentem alta importância global, sua influência varia signi-

ficativamente entre os locais analisados. O desvio-padrão da direção do vento (STWD) destaca-se como atributo dominante em várias estações, enquanto poluentes associados aos óxidos de nitrogênio e ao material particulado assumem maior relevância em contextos específicos, refletindo diferenças nos padrões locais de emissão e dispersão. Esses resultados evidenciam a heterogeneidade espacial do sistema monitorado e indicam que cada estação possui um conjunto particular de variáveis mais influentes, demonstrando o potencial das técnicas de XAI para revelar padrões locais de decisão e apoiar análises ambientais mais detalhadas.



(a) Contribuição relativa de variáveis. **(b) Importância relativa das variáveis.**

Figura 3. Análise de interpretabilidade do RF utilizando SHAP.

A análise em nível de estação evidencia a existência de regimes atmosféricos distintos, resultantes da interação entre fontes emissoras locais e condições meteorológicas. Na estação ACM, observa-se contribuição relevante de STWD e NO, indicando forte influência do tráfego veicular e sensibilidade à variabilidade do escoamento do vento em ambiente urbano complexo. A estação BR apresenta maior peso de NO, MP e STWD, sugerindo predominância de emissões veiculares e ressuspensão de partículas. Em CAB, destaca-se novamente a importância de STWD e NO, reforçando a influência combinada entre fontes locais de emissão e variabilidade direcional do vento. Já em CG, o impacto dominante de STWD, associado a contribuições secundárias de MP e RAIN, indica um regime fortemente controlado por processos de dispersão e remoção atmosférica. Na estação DT, além de STWD, observa-se maior relevância de WIND_SPEED e variáveis relacionadas à precipitação, sugerindo influência de processos de advecção e remoção úmida associados à circulação costeira. A estação IT apresenta um padrão mais equilibrado entre STWD, TEMP e NOx, indicando influência conjunta de estabilidade atmosférica e emissões locais. Em PI, destaca-se novamente a dominância de STWD acompanhada por NO e NOx, refletindo um ambiente urbano com forte influência de fontes veiculares. Por fim, a estação RV apresenta perfil mais distribuído entre variáveis meteorológicas e poluentes, com destaque para STWD, RAIN e O_3, sugerindo que o padrão provavelmente decorre da maior incidência de radiação solar e da influência de brisas marítimas na região costeira, condições favoráveis à formação fotoquímica de ozônio a partir de precursores transportados de áreas mais urbanizadas. Em conjunto, esses resultados indicam que cada estação possui uma assinatura atmosférica própria, determinada pelo balanço entre emissões locais, morfologia urbana e dinâmica meteorológica, evidenciando o potencial das técnicas explicáveis para compreender a variabilidade espacial da qualidade do ar.

4. Conclusão

Este trabalho apresentou uma abordagem baseada em modelos explicáveis de Machine Learning para a interpretação das relações entre poluentes atmosféricos e variáveis meteorológicas a partir de dados reais de monitoramento da cidade de Salvador, Bahia. O uso do modelo RF permitiu a classificação consistente das estações de monitoramento, enquanto a incorporação de técnicas de XAI possibilitou a análise detalhada dos fatores que influenciam o processo de decisão do modelo. Os resultados obtidos demonstraram que a metodologia adotada é capaz de capturar padrões atmosféricos relevantes, mantendo bom desempenho preditivo e fornecendo explicações interpretáveis associadas aos processos físicos e químicos do sistema analisado. A análise global e local baseada em SHAP evidenciou a predominância das variáveis meteorológicas na diferenciação entre as estações, com destaque para o desvio-padrão da direção do vento, a velocidade do vento e a precipitação, refletindo a importância dos processos de dispersão e remoção atmosférica em ambientes urbanos. Além disso, a análise por estação revelou forte heterogeneidade espacial, indicando que cada ponto de monitoramento apresenta uma assinatura atmosférica própria, determinada pelo balanço entre emissões locais, morfologia urbana e dinâmica meteorológica. Mudanças na frota veicular, no uso do solo ou na matriz industrial de Salvador ocorridas nos anos seguintes podem alterar as assinaturas atmosféricas aqui identificadas. Estudos futuros com dados mais recentes são necessários para avaliar a estabilidade temporal desses padrões. Esses achados reforçam o potencial das técnicas explicáveis para apoiar análises ambientais mais transparentes e contextualizadas, indo além da simples predição.

Agradecimentos

Os autores agradecem o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Agradecem também à CETREL S.A. pela disponibilização do conjunto de dados utilizado neste estudo.

Referências

Arunika, M., Saranya, S., Charulekha, S., Kabilarajan, S., and Kesavan, G. (2024). A survey on explainable ai using machine learning algorithms shap and lime. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Chakraborty, S., Misra, B., and Dey, N. (2024). Explainable artificial intelligence (xai) for air quality assessment. In *Design Studies and Intelligence Engineering*, pages 333–341. IOS Press.

Clifton, O., Paulot, F., Fiore, A., Horowitz, L., Correa, G., Baublitz, C., Fares, S., Goded, I., Goldstein, A., Gruening, C., et al. (2020). Influence of dynamic ozone dry deposition on ozone pollution. *Journal of Geophysical Research: Atmospheres*, 125(8):e2020JD032398.

Costa, E. L., Braga, T., Dias, L. A., Albuquerque, É. L. d., and Fernandes, M. A. (2022). Analysis of atmospheric pollutant data using self-organizing maps. *Sustainability*, 14(16):10369.

Costa, E. L., Braga, T., Dias, L. A., de Albuquerque, É. L., and Fernandes, M. A. (2024). Self-organizing maps applied to the analysis and identification of characteristics related

to air quality monitoring stations and its pollutants. *Neural Computing and Applications*, 36(19):11643–11657.

Das, A. and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Das, S., Agarwal, N., Venugopal, D., Sheldon, F. T., and Shiva, S. (2020). Taxonomy and survey of interpretable machine learning method. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 670–677. IEEE.

Emeç, M. and Yurtsever, M. (2025). A novel ensemble machine learning method for accurate air quality prediction. *International Journal of Environmental Science and Technology*, 22(1):459–476.

Méndez, M., Merayo, M. G., and Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, 56(9):10031–10066.

Monteiro, J. B. and Zanella, M. E. (2023). A metodologia estatística dos eventos extremos de precipitação: uma proposta autoral para análise de episódios pluviométricos diários. *Revista Brasileira de Climatologia*, 32:494–516.

Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., and Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in india. *Scientific reports*, 14(1):6795.

Parmar, A., Katariya, R., and Patel, V. (2018). A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and internet of things*, pages 758–763. Springer.

Rybarczyk, Y. and Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, 8(12):2570.

Shaik, A. B. and Srinivasan, S. (2018). A brief survey on random forest ensembles in classification model. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2*, pages 253–260. Springer.

Souza de Farias, G., Albuquerque, E., and Fernandes, M. (2026). Air Quality and Meteorological Dataset from Monitoring Stations in Salvador, Brazil, 2011–2016.

Stull, R. B. (2012). *An introduction to boundary layer meteorology*, volume 13. Springer Science & Business Media.

Tasioulis, T., Bagkis, E., Kassandros, T., and Karatzas, K. (2025). The quest for the best explanation: Comparing models and xai methods in air quality modeling tasks. *Applied Sciences*, 15(13):7390.

Wallace, J. M. and Hobbs, P. V. (2006). *Atmospheric science: an introductory survey*, volume 92. Elsevier.