Predição de Poluentes Atmosféricos via Métodos de Aprendizado de Máquina

Geovana R. P. Souza[1], Thiago I. A. Souza[1], Nicole S. Batista[1], Bruno V. Bertoncini[1], Danielo G. Gomes[2]

[1]Universidade Federal do Ceará
Núcleo de Pesquisa em Transportes e Meio Ambiente (TRAMA)

[2]Universidade Federal do Ceará
Grupo de Redes, Engenharia de Software e Sistemas (GREat)
Fortaleza – CE – Brasil

geovana21@alu.ufc.br, thiagoiachiley@ufc.br, nicolesouza@alu.ufc.br,

bruviber@det.ufc.br, danielo@ufc.br

Abstract. *Exposure to fine particulate matter ($PM_{2.5}$) poses a health risk in urban centers, demanding reliable forecasting systems. This paper proposes a predictive model based on machine learning applied to real-world data with 730,558 records collected by low-cost sensors in the city of Fortaleza, Brazil. We evaluated the performance of Random Forest, XGBoost, MLP, and SVR algorithms, following data preprocessing and calibration. The Random Forest model achieved the best performance, with an R^2 = 0.988 and an RMSE = 0.125. SHAP analysis identified PM_{10} e O_3 as the most relevant variables for prediction. The results suggest that artificial intelligence techniques can improve urban environmental monitoring and have strong potential to support data-driven e-Science platforms.*

Resumo. *A exposição a partículas finas em suspensão ($PM_{2,5}$) representa um risco à saúde em centros urbanos, exigindo sistemas de previsão confiáveis. Este artigo propõe um modelo preditivo baseado em aprendizado de máquina aplicado a dados reais com 730.558 registros coletados por sensores de baixo custo na cidade de Fortaleza/CE. Testamos os algoritmos Random Forest, XGBoost, MLP e SVR, após pré-processamento e calibração dos dados. O modelo Random Forest obteve o melhor desempenho, com R^2 de 0,988 e RMSE de 0,125. A análise SHAP revelou PM_{10} e O_3 como as variáveis mais relevantes para a predição. Os resultados indicam que técnicas de inteligência artificial podem melhorar o monitoramento ambiental urbano, com potencial para integrar plataformas de e-Ciência orientadas a dados.*

1. Introdução

A poluição do ar é uma ameaça crescente à saúde pública global, resultando na liberação de substâncias tóxicas no ambiente [Goudarzi et al. 2019]. Segundo relatório com dados da Organização das Nações Unidas (ONU), em 2016 quase 90% da população em grandes cidades respirou ar poluído, levando a 4,2 milhões de mortes devido à poluição [DESA 2023].

Com isso, sistemas de previsão da qualidade do ar são essenciais para estimar os níveis de qualidade do ar e fornecer informações rápidas, precisas e confiáveis para minimizar os efeitos adversos da poluição do ar na saúde humana e no meio ambiente. Sensores de baixo custo têm se destacado como alternativa viável aos sistemas tradicionais de monitoramento, por serem econômicos, compactos e fáceis de implantar, permitindo a formação de redes densas e gerando grandes volumes de dados espaço-temporais [Chojer et al. 2020].

A manutenção da qualidade do ar requer previsões rotineiras da poluição ambiental, a partir do monitoramento dessas densas redes de sensores. Existem muitos estudos na literatura para esse propósito, que utilizam modelos estatísticos e *machine learning* (aprendizado de máquina) para a predição [Rahman et al. 2024, Li and Sun 2021, Biancofiore et al. 2017]. Nos últimos anos, o desenvolvimento de modelos preditivos de qualidade do ar baseados em algoritmos de aprendizado de máquina surgiu como uma das direções de pesquisa para abordar as complexidades de grandes conjuntos de dados espaço-temporais de qualidade do ar, aprendendo a relação oculta dentro de dados históricos [Asgari et al. 2022].

Diante do contexto apresentado, este trabalho investiga a aplicação de técnicas de aprendizado de máquina na predição da qualidade do ar com base em dados reais coletados por 25 sensores com *hardware* de baixo custo distribuídos em uma cidade do nordeste brasileiro. Ao todo, 730.558 registros foram utilizados para prever os níveis de $PM_{2,5}$, com base em variáveis meteorológicas e poluentes atmosféricos. A abordagem proposta avalia quatro modelos de regressão e suas capacidades preditivas, sendo o modelo *Random Forest* o que alcançou o melhor desempenho na predição de ($PM_{2,5}$), com R2 = 0,988 e RMSE = 0,125. Esses resultados indicam o potencial de tais técnicas para aprimorar o monitoramento ambiental urbano.

2. Material e Método

A Figura 1 discrimina na forma de *timeline* o delineamento geral da proposta, desde o sensoriamento ambiental até a avaliação dos modelos preditivos de aprendizado de máquina utilizados. As etapas são explicadas no decorrer das subseções seguintes.

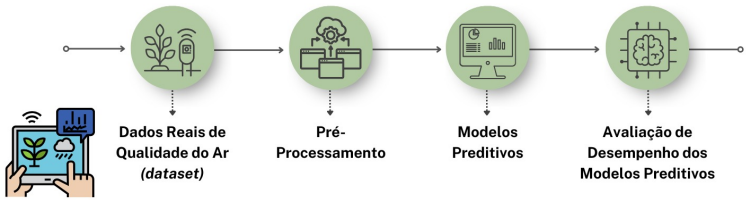

Figura 1. Timeline da modelagem de predição da qualidade do ar.

2.1. Dados Reais de Qualidade do Ar - Dataset

A coleta dos dados ocorreu por meio da rede de monitoramento da qualidade do ar composta por dispositivos denominados MoQA[1] (*Air Quality Monitor*). Implementado em maio de 2023, esse sistema é constituído por 25 sensores inteligentes de

[1]https://www.youtube.com/watch?v=KyhA3OIivfM

baixo custo, estrategicamente distribuídos em diferentes pontos da cidade de Fortaleza, abrangendo áreas escolares, corredores de tráfego, centros comerciais, unidades de saúde, espaços públicos e zonas industriais. Os dados coletados pelos dispositivos MoQA foram organizados em uma base de dados, previamente calibrado por meio de métodos de validação cruzada, baseados em abordagens testadas e validadas em uma pesquisa anterior [Alves 2023]. O conjunto de dados contém um total de 730.558 registros distribuídos em oito variáveis ambientais - Temperatura, Umidade, $PM_{2,5}$, PM_{10}, NO_2, CO_2 e O_3.

2.2. Pré-Processamento

Seja $\mathbf{X} = \{x_1, x_2, ...x_n\} \in R^{n \times d}$ o conjunto de atributos numéricos extraídos do monitoramento ambiental, onde n representa o número de amostras e d o número de atributos. O vetor de saída (variável alvo) é representado por $\mathbf{y} = \{y_1, y_2, ...y_n\} \in R^n$, sendo y_i correspondente à concentração de material particulado fino $PM_{2,5}$ na i-ésima amostra. Além disso, antes de iniciar o processamento matemático, todas as variáveis categóricas $\mathbf{x}_j \in R$, como colunas contendo *strings* ou identificadores sem relevância preditiva, foram descartadas para assegurar compatibilidade com os algoritmos supervisionados, que exigem entrada vetorial contínua.

Cada atributo $x_j \in R$, para $j = 1, 2, ..., d$, foi padronizado utilizando a normalização Z-score, representada por:

$$\widetilde{x}_j = \frac{x_j - \mu_j}{\sigma_j} \tag{1}$$

onde μ_j é a média do atributo j e σ_j é o desvio padrão.

2.3. Modelos Preditivos

2.3.1. Random Forest (RF)

Random Forest (RF) tem como princípio fundamental a construção de múltiplas árvores de decisão, selecionando amostras e características de forma aleatória, e então agregar os resultados por meio de votação ou média para obter a previsão final [Zou et al. 2025]. O algoritmo faz uso da técnica *bootstrap sampling*, na qual são geradas várias amostras do conjunto de dados de treinamento com reposição. Para cada uma dessas amostras, uma árvore de decisão individual é construída utilizando um subconjunto aleatório dos atributos disponíveis. Esse processo garante diversidade entre as árvores, tornando o modelo mais generalizável à medida que o número de T árvores aumenta:

$$H(\mathbf{x}) = \frac{1}{T} \sum_{i=1}^{T} h_i(\mathbf{x}) \tag{2}$$

em que $h_i(x)$ é a saída da i-ésima árvore de regressão h_i para a amostra x.

2.3.2. XGBoost (*eXtreme Gradient Boosting*)

O método *XGBoost* tem como suas características mais importantes, a capacidade de lidar com dados ausentes, evitar *overfitting*, alcançar alto poder preditivo e execução

rápida [Cengil 2025]. O algoritmo foi desenvolvido com base no conceito de *boosting*, no qual cada nova árvore é treinada para corrigir os erros da árvore anterior. Essa correção de erros ocorre minimizando iterativamente a função de perda, dada pela equação:

$$L(t) = \sum_{i=1}^{n} l(y_i, y^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3}$$

onde l é a função de perda, t representa a iteração atual, $t-1$ as iterações anteriores e $\Omega(f_t)$ é um termo de regularização que auxilia na redução do *overfitting*.

2.3.3. Multilayer Perceptron (MLP)

O *Multilayer Perceptron* (MLP) é um algoritmo de aprendizado de máquina inspirado no funcionamento do cérebro humano, sendo composto por, no mínimo, três camadas: a camada de entrada, responsável por receber os dados; uma ou mais camadas ocultas, que realizam o processamento intermediário; e a camada de saída, que entrega a predição final. Cada camada é formada por unidades chamadas *perceptrons*, que constituem os blocos fundamentais da rede. A capacidade de aprender padrões complexos e não lineares nos dados se deve, em grande parte, ao uso de funções de ativação nas camadas ocultas. Considerando i como a i-ésima camada da rede e j como a j-ésima unidade oculta da camada, tem-se:

$$z_j^{[i]} = W_j^{[i]^T} \cdot x + b_j^{[i]} \tag{4}$$

em que x representa o vetor de entrada do neurônio, $W_j^{[i]}$ é o vetor de pesos, $b_j^{[i]}$ é o termo de viés, e $z_j^{[i]}$ é a saída linear (pré-ativação) do neurônio [Jairi et al. 2024].

2.3.4. SVR (*Support Vector Regression*)

O SVR (*Support Vector Regression*) é uma variação do SVM (*Support Vector Machine*), utilizada para tarefas de regressão. Seu principal objetivo é encontrar uma função que melhor estime o valor de saída com base nos valores de entrada. O mapeamento de entrada-saída para casos não lineares [Galli et al. 2018], utilizando uma função kernel, assume a forma:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{5}$$

onde, α_i e α_i^* são os multiplicadores de Lagrange;$K(x_i, x)$ é a função kernel, que transforma os dados de entrada para um espaço de características de maior dimensão; e b é o viés do modelo.

2.4. Avaliação de Desempenho dos Modelos Preditivos

Os dados foram divididos em 80% para treino e 20% para teste, proporção constantemente utilizada em modelagens preditivas, sendo normalizados no intervalo de 0 a 1. Para os algoritmos *Random Forest* e *XGBoost*, foram escolhidas 100 árvores de regressão. Além disso, o modelo *XGBoost* contou com uma taxa de aprendizado de 0,1 e uma profundidade máxima das árvores de 6.

As métricas adotadas para avaliar o desempenho dos modelos de aprendizado de máquina foram o coeficiente de determinação (R^2), dada pela expressão a seguir:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

e a raiz do erro quadrático médio (RMSE, do inglês *Root Mean Squared Error*), dada pela equação a seguir:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

onde y_i corresponde ao valor real da i-ésima amostra, $\hat{y}_i$ é o valor previsto da i-ésima amostra e $\bar{y}$ representa a média das observações.

3. Resultados e Discussões

Este artigo usou abordagem de aprendizado de máquina para prever a qualidade do ar usando um conjunto de dados real. Para esse propósito, todo esse estudo foi realizado usando Python (v.3.10), uma linguagem de programação, em um computador de processador i3, placa GPU e 12 GB de RAM. Nessa perspectiva, esta seção se concentra na predição de concentrações do poluente atmosférico $PM_{2,5}$ na capital de Fortaleza/CE desenvolvida com os modelos de aprendizado de máquina RF, XGBoost, MLP e SVR. O $PM_{2,5}$ foi escolhido como variável alvo por seu potencial danoso à saúde, uma vez que pode facilmente entrar nos pulmões levando vírus, o que representa sérios riscos à saúde, particularmente para populações vulneráveis, como crianças e idosos [Lakra et al. 2025].

Primeiro, a variação e variabilidade dos diferentes poluentes atmosféricos é apresentada no *violin plot* (vide Figura 2), que exibe a distribuição das características finais dos dados. Para aumentar a clareza dos registros, o logaritmo dos dados coletados foi calculado na etapa de pré-processamento.

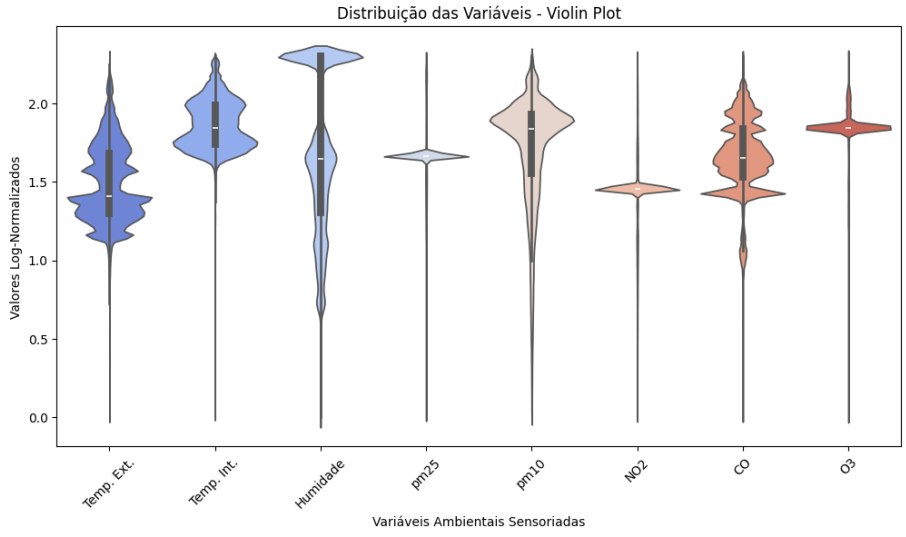

Figura 2. Diagrama violin plot comparando os poluentes atmosféricos sensoriados.

Segundo, o desempenho dos modelos implementados é apresentado na Tabela 1, a seguir. A relação entre o $PM_{2,5}$ previsto e os níveis reais de $PM_{2,5}$ foi descrita com valores de R^2 e RSME nos modelos RF, XGBoost, MLP e SVR, com os melhores valores de R^2 e RSME de 0,988 e 0,125, respectivamente, para RF, e valores de R^2 e RSME de 0.924 e 0.318, respectivamente, para o MLP. O modelo RF forneceu o menor valor de RMSE. Em outras palavras, ele mostrou o desempenho de previsão mais bem-sucedido, enquanto o modelo SVR teve a maior taxa de erro, particularmente, RMSE (6,053). A maior determinação dos valores de R^2 explica que o modelo RF é mais eficiente e melhor do que o modelo MLP na previsão de $PM_{2,5}$. O melhor desempenho do modelo RF se deve à superioridade bem conhecida em problemas e relações não lineares e modelagens complexas com $PM_{2.5}$ [Kawichai et al. 2025].

Tabela 1. Desempenho dos modelos preditivos

Modelo	R^2	RMSE
Random Forest	**0.9880**	**0.1250**
MLP	0.9240	0.3180
XGBoost	0.9081	0.3510
SVR	-26.499	6.053

Com base nos resultados obtidos, uma simulação do desempenho de saída dos modelos RF, XGBoost, MLP e SVR versus valores-alvo de previsão dos níveis de $PM_{2,5}$ é mostrada na Figura 3. Cruzando com o resultado apresentado na Tabela 2 para o maior valor do coeficiente (R^2) do RF, a Figura 3a exibe uma correlação relativamente alta entre os valores de saída e alvo para a previsão da concentração de $PM_{2,5}$ na cidade de Fortaleza/CE. A Figura 3b exibe a segunda melhor correlação, para o modelo XGBoost. Por fim, a Figura 3d apresenta o comportamento da correlação para o modelo que apresentou o maior erro e o menor R^2, o SVR.

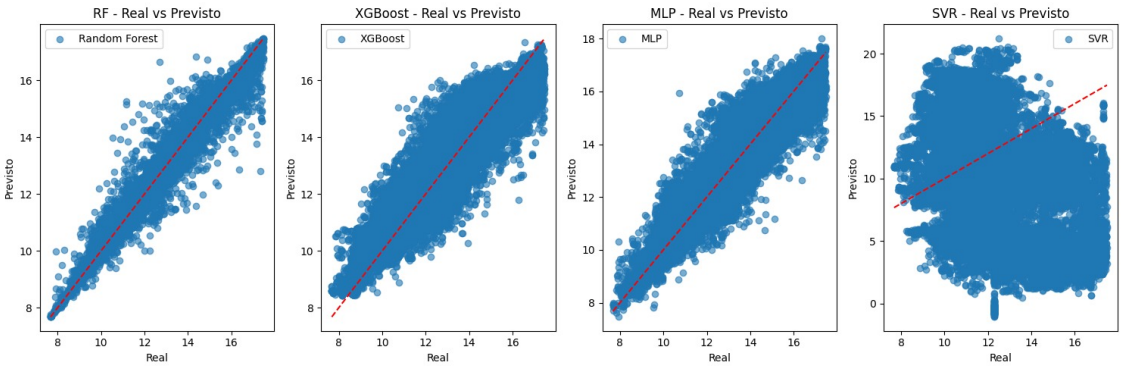

Figura 3. Diagramas de dispersão de concentrações observadas e previstas do poluente $PM_{2,5}$ dos modelos Random Forest, XGBoost, MLP e SVR.

Finalmente, a Figura 4 mostra o efeito das variáveis na concentração prevista de $PM_{2,5}$ usando o modelo de maior R^2 e menor erro, particularmente, o modelo RF. Esses efeitos foram medidos usando gráficos SHAP, que fornecem *insights* sobre como cada uma das variáveis contribui para a saída do modelo [Deveer and Minet 2025]. Valores SHAP positivos indicam que uma variável contribui para aumentar o valor previsto (como

é o caso das variáveis PM_{10} e O_3), enquanto valores SHAP negativos significam que a variável contribui para diminuir o valor preditivo do modelo (como é o caso da temperatura externa). Da mesma forma, os pontos nos gráficos representam o valor de atribuição de cada variável e são coloridos com base nos valores da variável - altos impactos são coloridos em vermelho (novamente as variáveis PM_{10} e O_3) e baixos impactos são coloridos em azul (novamente a temperatura externa). Quando os pontos são empilhados verticalmente (ou próximos de zero), significa que a variável não altera significativamente o resultado do modelo para diferentes observações (como é o caso das variáveis humidade, NO_2, CO e temperatura interna).

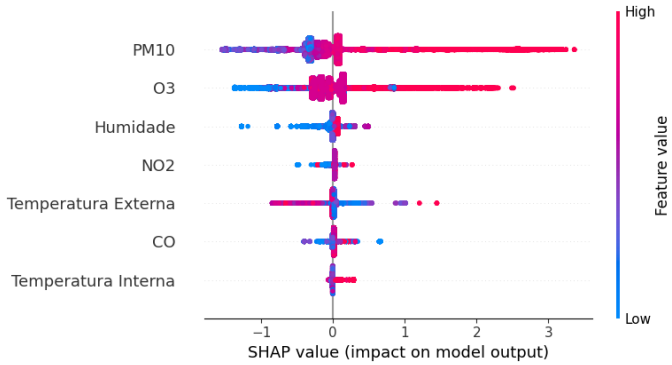

Figura 4. SHAP do modelo RF para concentrações de $PM_{2,5}$.

4. Conclusão

Este artigo mostrou que é possível prever concentrações de $PM_{2,5}$ com alta precisão utilizando sensores de baixo custo e técnicas de aprendizado de máquina. Dentre os modelos avaliados, o *Random Forest* obteve os melhores resultados. A análise SHAP permitiu compreender o peso relativo das variáveis, destacando PM_{10} e O_3 como principais influenciadores das previsões. Esses achados reforçam o papel da inteligência artificial como ferramenta de apoio à ciência ambiental e à gestão urbana. Como continuidade, propõe-se expandir a modelagem para outros poluentes e explorar arquiteturas mais complexas, como redes neurais profundas e modelos híbridos em ambientes computacionais distribuídos.

Agradecimentos

Danielo G. Gomes agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de produtividade (processo 311845/2022-3).

Referências

Alves, M. L. A. (2023). Repoair: Desenvolvimento do repositório de dados para monitoramento da qualidade do ar. Trabalho de conclusão de curso (graduação em engenharia civil), Universidade Federal do Ceará, Fortaleza. Orientador: Prof. Dr. Bruno Vieira Bertoncini.

Asgari, M., Yang, W., and Farnaghi, M. (2022). Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework. *Environmental Technology Innovation*, 27:102776.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Tommaso, S. D., Colangeli, C., Rosatelli, G., and Carlo, P. D. (2017). Recursive neural network model for analysis and forecast of pm10 and pm2.5. *Atmospheric Pollution Research*, 8:652–659.

Cengil, E. (2025). The power of machine learning methods and pso in air quality prediction. *Applied Sciences*, 15:2546.

Chojer, H., Branco, P., Martins, F., Alvim-Ferraz, M., and Sousa, S. (2020). Development of low-cost indoor air quality monitoring devices: Recent advancements. *Science of The Total Environment*, 727:138385.

DESA, U. (2023). The sustainable development goals report 2023: Special edition - july 2023. Technical report, Disponível em: https://unstats.un.org/sdgs/repor t/2023/. Acesso em: 11 març 2025.

Deveer, L. and Minet, L. (2025). Real-time air quality prediction using traffic videos and machine learning. *Transportation Research Part D*, 142:104688.

Galli, L., Galvan, G., Sciandrone, M., Cantù, M., and Tomaselli, G. (2018). Machine learning methods for short-term bid forecasting in the renewable energy market: A case study in italy. *Windy Energy*, 21.

Goudarzi, G., Shirmardi, M., Naimabadi, A., Ghadiri, A., and Sajedifar, J. (2019). Chemical and organic characteristics of pm2.5 particles and their in-vitro cytotoxic effects on lung cells: The middle east dust storms in ahvaz, iran. *Science of The Total Environment*, 655:434–445.

Jairi, I., Ben-Othman, S., Canivet, L., and Zgaya-Biau, H. (2024). Enhancing air pollution prediction: A neural transfer learning approach across different air pollutants. *Environmental Technology Innovation*, 36.

Kawichai, S., Sripan, P., Rerkasem, A., Rerkasem, K., and Srisukkham, W. (2025). Long-term retrospective predicted concentration of pm2.5 in upper northern thailand using machine learning models. *Toxics*, 13:170.

Lakra, A. R., Gautam, S., Samuel, C., and Blaga, R. (2025). College bus commuter exposures to air pollutants in indian city: The urban-rural transportation exposure study. *Geosystems and Geoenvironment*, 4:100346.

Li, Y. and Sun, Y. (2021). Modeling and predicting city-level co2 emissions using open access data and machine learning. *Environmental Science and Pollution Research*, 28:19260–19271.

Rahman, M., Nayeem, E. H., Ahmed, S., Tanha, K. A., Sakib, S. A., Hafiz, K. M. M. U. ., and Babu, H. (2024). Airnet: predictive machine learning model for air quality forecasting using web interface. *Environmental Systems Research*, 13:1–19.

Zou, Y., Tian, H., Huang, Z., Liu, L., Xuan, Y., Dai, J., and Nie, L. (2025). Study on prediction models of oxygenated components content in biomass pyrolysis oil based on neural networks and random forests. *Biomass and Bioenergy*, 193:107601.