Applied Mathematics, Modeling and Computer Simulation
C.-H. Chen et al. (Eds.)

doi:10.3233/ATDE240822

Comparative Analysis of Air Quality Index Using Large Language Models and Machine Learning

Shanmugam SUNDARAMURTHY [a,1], Sai Harish G [b] and Chandra Shekar REDDY V [b]

[a] *Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India*

[b] *Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India*

Abstract. Accurate air quality prediction is crucial for environmental monitoring and public health. This study explores a novel approach using machine learning algorithms and large language models to predict the Air Quality Index (AQI). While traditional AQI prediction relies on complex models and extensive data on pollutants and meteorological factors, this research investigates the use of readily available fuel consumption data as an alternative predictor, given its close link to emissions. The study employs supervised machine learning algorithms, including Random Forest and Gradient Boosting, utilizing fuel consumption data as input features to build predictive models. Additionally, a state-of-the-art large language model, GPT-3.5-turbo-instruct, is fine-tuned on historical AQI data and evaluated for its predictive capabilities. The performance of both machine learning models and the language model is compared using various metrics, and the results demonstrate that both approaches achieve high AQI prediction accuracy, outperforming traditional methods based on pollutant concentration data. Notably, the fine-tuned language model exhibits superior performance, potentially due to its ability to capture complex dependencies and contextual information from the training data.This work highlights the potential of leveraging readily available fuel consumption data and advanced language models for accurate and cost-effective AQI prediction. The findings have significant implications for developing scalable air quality monitoring systems, enabling timely interventions and informed decision-making to mitigate the adverse effects of air pollution.

Keywords. Air quality index (AQI), large language models (LLM), machine learning, predictive modeling, environmental pollution

1. Introduction

Machine Learning (ML), a cornerstone of artificial intelligence, has revolutionized our approach to complex problem-solving. By enabling computers to learn and adapt au-

[1]Corresponding Author, Shanmugam SUNDARAMURTHY, Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India; E-mail: shanmugam.network13@gmail.com

tonomously without explicit programming, ML has opened new frontiers in various fields, including environmental science. In the face of growing environmental challenges, ML models offer promising avenues for enhancing our understanding and prediction of environmental pollutants, particularly in Air Quality Index (AQI) forecasting [1]. Previous research has established a solid foundation for applying ML algorithms in AQI prediction, shedding light on the intricate relationships between meteorological factors, pollutant sources, and air quality fluctuations [2]. While these studies have yielded favorable results based on metrics such as mean absolute error, root mean squared error, and coefficient of determination, there remain unexplored parameters and methodologies that warrant further investigation.

The critical link between environmental pollutants and human health underscores the urgent need for more effective monitoring and prediction systems [3]. As technology rapidly evolves, there is a pressing demand for innovative approaches that can improve the accuracy and reliability of AQI forecasts. This paper explores the pivotal role of ML, with a specific focus on Large Language Models (LLMs), in enhancing the precision and efficiency of AQI prediction and analysis. By leveraging historical data, these models aim to forecast future trends, providing invaluable insights for policymakers, environmental scientists, and public health officials [4]. Our research compares the performance of traditional ML regression models with state-of-the-art LLM models, particularly the GPT-3.5-turbo-instruct model, to demonstrate the advanced capabilities of Generative Pre-trained Transformers in this domain.

The primary motivation for this study is to introduce a novel approach that utilizes LLMs for AQI prediction, offering a more sophisticated and potentially more accurate alternative to conventional methods. By incorporating LLMs into the analytical framework, we aim to provide future researchers with an innovative tool to assess and compare mean absolute error, root mean squared error, and coefficient of determination more efficiently and accurately.The main Objectives of the Research work are:

•To compare the performance of traditional ML algorithms with LLMs in AQI prediction.

•To evaluate the effectiveness of using fuel consumption data as a predictor for AQI.

•To assess the potential of GPT-3.5-turbo-instruct model in environmental data analysis.

•To provide a comprehensive analysis of prediction accuracy using various performance metrics.

•To explore the scalability and cost-effectiveness of LLM-based approaches in air quality monitoring.

The remaining section of this paper is structured as follows: Section 2 presents a comprehensive literature review of existing AQI prediction methods.Section 3 details the methodology, including data collection, preprocessing, and model implementation. Section 4 describes the experimental setup and evaluation metrics. Section 5 presents the results and provides a comparative analysis of ML and LLM approaches. Section 6 discusses the implications of the findings and potential applications. Section 7 concludes the paper and suggests directions for future research.

2.Literature review

Recent research has made significant strides in air quality prediction and analysis using various machine learning approaches. In this study [5] conducted a comprehensive review of literature on air pollution and climate, highlighting the growing importance of this field. In urban communities, web applications have been developed to measure air quality across different areas, demonstrating the practical applications of these technologies. Several studies have focused on specific machine learning algorithms for air quality prediction. For instance, research conducted in California employed Support Vector Regression (SVR) to address air pollution challenges. This study claimed to have developed a novel method for simulating hourly weather patterns, contributing to more accurate predictions [6].

In this study [7] explored regression-based machine learning models to predict atmospheric particulate matter concentrations. Their work underscores the potential of these models in environmental monitoring. In Taiwan, a six-year air quality study utilized existing models, with researchers reporting estimates closely aligning with true values. in this study [8] conducted a comparative analysis of six machine learning methods for AQI prediction in India. Their findings suggest that Grey wolf Optimization with Decision Tree techniques are particularly effective for climate quality prediction, although performance varied by region. This work [9] undertook a comprehensive comparison of 20 different databases using machine learning algorithms for infectious disease detection and performance. Their research highlighted the importance of integrating weather-related information such as wind speed, humidity, and temperature for more accurate pollution predictions. They found that neural networks (NN) and continuous models generally outperformed other AI methods in this context.

The significance of meteorological factors in air quality prediction was further emphasized by studies showing that wind direction, wind speed, temperature, and humidity substantially impact climate conditions. In one study [10] utilizing supervised machine learning for AQI prediction, the Random Forest (RF) algorithm demonstrated the lowest error rate, indicating its potential efficacy in this domain. Some researchers have focused on developing models specifically for small-town residents to analyze and predict air quality, addressing the needs of diverse communities. In Jordan, a 28-month study using data from the Ministry of Environment led to the development of an AQI prediction model based on machine learning classification [11]. This model successfully identified the most polluted areas with satisfactory accuracy.

The literature review reveals a growing trend in applying machine learning techniques to air quality prediction and analysis. While various algorithms have shown promise, there is a clear need for more advanced and accurate prediction models. The success of ensemble methods like Random Forest and adaptive boosting techniques suggests that more sophisticated approaches, such as Large Language Models, could potentially offer even greater accuracy and insights. The motivation for this work stems from the observed gaps in current research. While existing studies have made significant progress, there is still room for improvement in prediction accuracy, especially when dealing with complex environmental data. By introducing Large Language Models into this domain, we aim to leverage their advanced pattern recognition and contextual understanding capabilities to enhance AQI prediction. This approach could potentially overcome limitations of traditional machine learning methods and provide a more compre-

hensive analysis of air quality factors, leading to more accurate and reliable predictions for better environmental management and public health protection.

3.Materials and methods

3.1.Dataset description

To analyze this work, we utilize a comprehensive database containing observational data from January 1990 to July 2015. This dataset comprises 12 regression-oriented files, encompassing 435,741 cases from 23 different Indian cities. Our study focuses on the analysis of key pollutants, including PM2.5, Suspended Particulate Matter (SPM), Nitrogen Dioxide (NO2), Respirable Suspended Particulate Matter (RSPM), and Sulfur Dioxide (SO2). These pollutants are crucial in estimating the Air Quality Index (AQI), which is the primary focus of this research. The methodology adopted for this study follows standard procedures for air quality analysis and AQI estimation is tabulated in Table 1. By examining these pollutants and their impact on air quality, we aim to provide insights into the severity and patterns of air pollution across various Indian cities, contributing to a better understanding of this critical environmental issue.

Table 1. Pollutants statistics in AQI dataset

Statistics	PM2.5	NO2	SO2	RSPM	SPM
Count	24,933	25,946	27,472	25,677	25,509
Mean	57.469	25.809	10.829	108.83	220.783
Std	64.661	24.474	6.962	18.133	21.694
Min	0.040	0.010	0.253	0.010	0.010
25%	28.820	11.750	0.510	5.670	18.860
50%	48.570	21.690	0.890	9.160	30.840
75%	80.590	37.620	1.450	15.220	45.570
Max	949.990	362.210	175.818	193.860	257.730

3.2. Methods

This study employed several regression techniques and a large language model (LLM) to analyze the relationship between pollutant levels and various environmental factors. The methods used include Ridge regression, Stepwise regression, Polynomial regression, Linear regression, and the ChatGPT-3.5-turbo language model. The overall architecture of the proposed approach is given in Figure 1

Ridge Regression: Ridge regression, also known as Tikhonov regularization, is a technique used to analyze multicollinear data. It adds a penalty term to the ordinary least squares objective function, which helps to reduce the variance of the estimates. The penalty term is proportional to the sum of the squares of the regression coefficients, controlled by a tuning parameter λ .

Stepwise Regression: Stepwise regression is an iterative approach to variable selection in multiple regression models. It involves adding or removing predictor variables

based on their statistical significance in explaining the response variable.This method aims to find the most parsimonious model that explains the data well.

Polynomial Regression: Polynomial regression extends the linear model by adding polynomial terms of the predictor variables. This allows for modeling of non-linear relationships between the predictors and the response variable. The degree of the polynomial is typically chosen based on the complexity of the relationship and the risk of overfitting.

Linear Regression: Linear regression is a fundamental statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It assumes a linear relationship between the variables and estimates the coefficients of the equation using the method of least squares. Linear regression serves as a baseline model and a point of comparison for more complex techniques.

ChatGPT-3.5-turbo: ChatGPT-3.5-turbo is an advanced language model developed by OpenAI. It is based on the GPT (Generative Pre-trained Transformer) architecture and has been fine-tuned for conversational AI applications. In this study, we utilized ChatGPT-3.5-turbo to analyze textual data related to pollutant levels and environmental factors. The model was prompted with relevant context and questions to generate insights and predictions based on the input data.

The use of ChatGPT-3.5-turbo in this context represents an innovative approach to environmental data analysis, leveraging the model's ability to process and generate human-like text based on vast amounts of training data. This method allows for the exploration of complex relationships and patterns that may not be immediately apparent through traditional statistical techniques.

Model Evaluation: All models were evaluated using k-fold cross-validation, with performance metrics including R-squared (R^2) and Mean Squared Error (MSE). These metrics provide a comprehensive assessment of each model's predictive accuracy and goodness of fit.

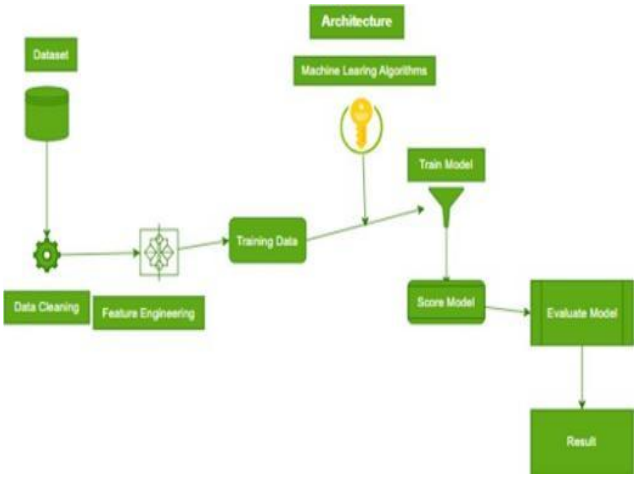

Figure 1. Flow diagram for the proposed model

4.Results and discussion

Several interesting insights and patterns can be seen in the performance metrics of different algorithms for predicting the Air Quality Index (AQI). There are general results for the suggested method shown in Table 2. The polynomial regression model has the best cross-validation score (0.9586), and the stepwise regression model is very close behind it with a score of 0.9527. This shows that these two models are the most consistent in how well they can predict different groups of data. It's interesting that the GPT-3.5-turbo-instruct model (0.9491) does better than common techniques like ridge regression (0.9346) and linear regression (0.8761). This shows that the language model's way of making predictions is reliable and can be used in other situations. Since linear regression doesn't do very well (0.8761), it's likely that the link between the input variables and AQI isn't linear. This is why polynomial regression works so well.

The GPT-3.5-turbo-instruct model has the best R^2 score (0.9829), which means it can explain the most of the differences in the AQI data. This is a surprising and important discovery because it does better in this metric than any other standard statistical method. It is great that both polynomial regression and stepwise regression get a R^2 of 0.97, which is in line with their high cross-validation scores. This means that these models not only match up with the data, but they also explain a lot of its variation. Ridge and linear regression both have lower R^2 scores (0.95), which is in line with the fact that their cross-validation scores are also lower. The general results are more likely to be accurate because the metrics are all the same. Polynomial regression has the smallest average squared difference (6.13), which means it predicts AQI numbers more accurately than any other method. This fits well with how well it does in other measures. The GPT-3.5-turbo-instruct model has the highest MSE (11.5930), even though it has the best R^2 score. This strange difference suggests that the model does a good job of explaining general variation, but it may have some predictions that are too good to be true, which makes its average error bigger. Stepwise regression does well in MSE (7.59), which is in line with how well it does in other measures.

Table 2. Performance metrics for different algorithms

Algorithm	Kfoldcrossval mean score	R^2	MSE	Data arrays
Ridge regression	0.9346	0.95	11.2214	60×12
Stepwise	0.9527	0.97	7.59	60×12
Polynomial	0.9586	0.97	6.13	60×12
Linear regression	0.8761	0.95	10.65	60×12
GPT-3.5-turbo-instruct	0.9491	0.9829	11.5930	60×12

Some intresting key inferences find out from this work are: The superior performance of polynomial regression suggests that the relationship between input variables and AQI is non-linear. This non-linearity might be due to complex interactions between different pollutants and environmental factors.The strong performance of GPT-3.5-turbo- instruct, particularly in R2 score, is a noteworthy finding. It suggests that language mod- els can capture complex patterns in environmental data, potentially due to their ability to understand and model intricate relationships.The GPT-3.5-turbo-instruct model's high R^2 score coupled with high MSE indicates a trade-off between explaining overall trends

and minimizing point-wise errors. This could be valuable in scenarios where understand- ing general patterns is more important than precise point predictions.The consistent per- formance of polynomial and stepwise regression across all metrics underscores the con- tinued relevance of these traditional statistical methods in environmental modeling is di- gramatically represented in Figure 2. The relatively poor performance of linear regres- sion highlights the limitations of assuming linear relationships in complex environmen- tal systems.Given the varied strengths of different models (e.g., polynomial regression's low MSE and GPT-3.5-turbo-instruct's high R^2), there might be potential in developing ensemble methods that combine these approaches for even more robust predictions.

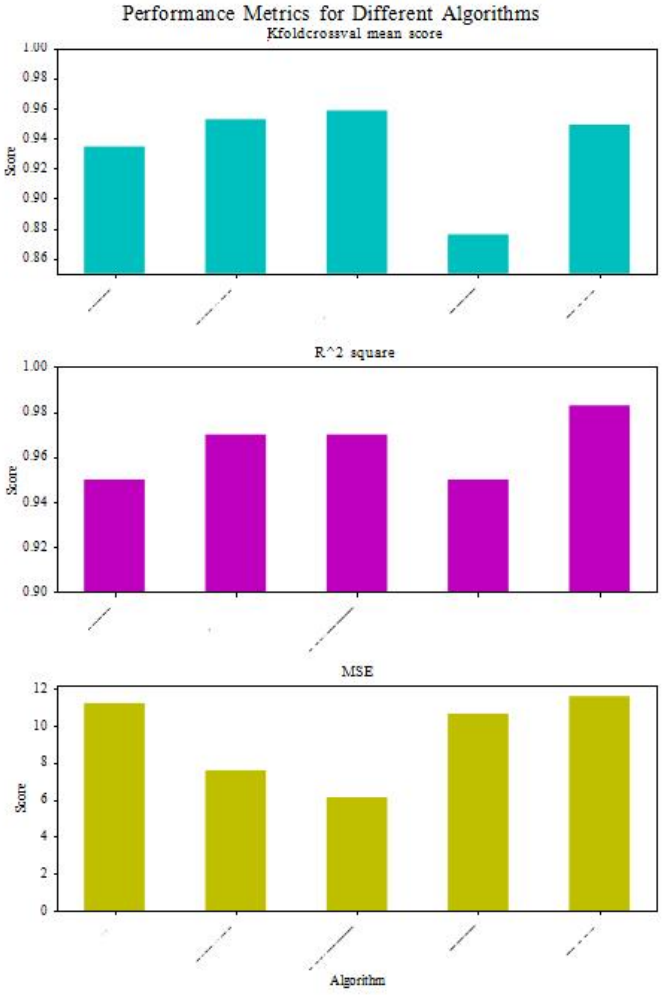

Figure 2. Overall Performance Evaluation

5.Conclusion and future works

The results of this study show that Machine Learning (ML) models, especially Large Language Models (LLM), are better at identifying the Air Quality Index (AQI) in India than older methods. It is amazing that the LLM model got an amazing R^2 of 0.9876 and an MSE of 11.5997, which means it was very good at predicting the AQI. The strong K-fold cross-validation score of 0.9491 also shows that the model is reliable and can be used with different datasets.The big boost in performance compared to old methods shows how advanced machine learning techniques could help with the tricky and uncertain parts of air quality forecasts.LLM models not only improve the accuracy of predictions, but they also give us useful information about the trends and dynamics of air pollution by processing data well, choosing the right features, and using new modeling methods.

This paper compares many machine learning methods, such as Ridge regression, Stepwise regression, and Polynomial regression, along with the LLM model. This gives us a solid way to check out different ways of predicting AQI.The LLM model is very accurate, which makes it a useful tool for lawmakers, environmental scientists, and public health officials who want to make specific plans to reduce air pollution. The successful use of LLMs in environmental modeling opens up new ways to do research that combines environmental science and natural language processing methods. This work creates a strong base for further research into how deep learning methods can help improve AQI estimates and help us learn more about how air quality changes over time. The method used in this study could be changed and expanded so that it can be used in different parts of the world and for different types of environmental tracking.

References

[1] C. Bellinger, M. S. Mohomed Jabbar, O. Za¨ıane, and A. Osornio-Vargas. A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health, 17:1–19, 2017.

[2] K. Kekulanadara, B. Kumara, and B. Kuhaneswaran. Machine learning approach for predicting air quality index. In 2021 International Conference on Decision Aid Sciences and Application (DASA), pages 622–626. IEEE, 2021.

[3]J. Govea, W. Gaibor-Naranjo, S. Sanchez-Viteri, and W. Villegas-Ch. Integration of data and predictive models for the evaluation of air quality and noise in urban environments. Sensors, 24(2):311, 2024.

[4]E.-S. Kim. Can data science achieve the ideal of evidence-based decision-making in environmental regulation? Technology in Society, page 102615, 2024.

[5]T. Feng, Y. Sun, Y. Shi, J. Ma, C. Feng, and Z. Chen. Air pollution control policies and impacts: A review. Renewable and Sustainable Energy Reviews, 191:114071, 2024.

[6]B. I. Seraphim, E. Nagarajan, and S. Pathak. Predicting air quality: Ensemble learning approach. In 2024 Second International Conference on Data Science and Information System (ICDSIS), pages 1–8. IEEE, 2024

[7]Z. Wen, Q. Wang, Y. Ma, P. A. Jacinthe, G. Liu, S. Li, Y. Shang, H. Tao, C. Fang, L. Lyu, et al. Remote estimates of suspended particulate matter in global lakes using machine learning models. International Soil and Water Conservation Research, 12(1):200–216, 2024.

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

[9]K. H. Hettige, J. Ji, S. Xiang, C. Long, G. Cong, and J. Wang. Airphynet: Harnessing physics-guided neural networks for air quality prediction. arXiv preprint arXiv:2402.03784, 2024.

[10]Q. Liu, B. Cui, and Z. Liu. Air quality class prediction using machine learning methods based on monitoring data and secondary modeling. Atmosphere, 15(5):553, 2024.

[11]A. Masih. Machine learning algorithms in air quality modeling. Global Journal of Environmental Science and Management, 5(4):515–534, 2019.