AQI Predictor

Ayush Sharma[#1], Dhairya Rajput[#2], Ankit[#3], Sandeep Kumar[#4], Suraj Bhatnagar[#5]

[#]*Computer Science and Engineering (Data Science), Meerut Institute of Engineering and Technology (MIET), Meerut, India*

[1] ayushpsharma88099@gmail.com

[s2]drajput9760@gmail.com

[3]ankitk908432@gmail.com

[4]2002sandeep145@gmail.com

[5]suraj.bhatnagar@miet.ac.in

***Abstract*— Air pollution is a significant environmental problem that affects urban areas around the world, and it has negative effects on both the environment and the public health| Rapid urbanization and industrialization in cities like Meerut, India, have resulted in poor air quality. It is necessary to take proactive action to address this problem| Through the use of data-driven techniques like web scraping, data analysis, machine learning algorithms, and data visualization, this project aims to create an Air Quality Index (AQI) predictor for Meerut city| The proposed system collects historical air quality data, preprocesses it, uses different algorithms to train regression models, and uses Tableau to show AQI trends and insights| Average Absolute Error (MAE) and Root Mean Squared Error (RMSE) are performance metrics used to assess the system's efficiency and accuracy.**

The experiment results show that the AQI predictor is good at predicting air quality and encouraging action to reduce pollution| Environmental sustainability and a healthier urban environment in Meerut and beyond are both promoted by this project.

***Keywords*— Web scraping, data preprocessing, regression algorithms, Linear Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, XGBoost Regression.**

1. INTRODUCTION

India, is facing serious issues of air pollution, which is a major issue affecting urban areas around the world| By creating an Air Quality Index (AQI) predictor for Meerut, this initiative aims to combat this issue| The system aims to provide useful insights for reducing air pollution and improving public health in the area by using advanced technologies and data analysis methods.

1.1. Background History

The growing concern about air pollution in Meerut has prompted the development of an Air Quality Index (AQI) predictor| Meerut, like many other cities, is facing problems due to accelerated urbanization and industrialization, which leads to poor air quality and bad health effects| The project aims to provide accurate, timely air quality forecasts by using data-driven approaches and technological advancements, inspired by successful AQI prediction models that are used worldwide| The AQI predictor aims to give stakeholders actionable insights to reduce air pollution and improve Meerut's cities by analyzing historical data and using machine learning algorithm.http://www.cs.utsa.edu/~shxu/socs/

Supported Technologies and Algorithms

The Air Quality Index (AQI) predictor uses web scraping to get historical data| Machine learning algorithms such as Linear Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, and XG Boost Regression model find AQI values, while data analysis techniques find patterns and correlations| To effectively present insights to stakeholders, data visualization tools like Tableau are then used.

2. LITERATURE REVIEW

The creation of Air Quality Index (AQI) prediction models has been the subject of numerous studies in recent years, which is indicative of the increasing significance of proactive air quality management in urban areas across the globe. Liu et al.'s (2019) investigation on the effectiveness of machine learning algorithms in AQI prediction is noteworthy among these works [1]. By means of their study, they were able to exhibit the potential of sophisticated algorithms in precisely anticipating AQI values by proving the superior performance of ensemble approaches such as Random Forest and XGBoost over conventional regression procedures.

Similar to this, Zhang et al. (2020) made significant contributions to our understanding of how data mining methods can be used to forecast AQI [2]. Their research highlighted how important feature selection and model assessment are to getting accurate forecasts. Through the utilization of data mining techniques, scholars can derive significant patterns and connections from intricate datasets on air quality, augmenting the prognostic potential of AQI models.

Furthermore, research like those done by Wang et al. (2018) and Sharma et al. (2021) has shown how crucial it is to include meteorological aspects in AQI prediction models [3]. The impact of meteorological factors, including temperature, humidity, and wind speed, on the dynamics of air quality was emphasized by these researchers. Prediction accuracy is improved and a more thorough understanding of air pollution patterns is made possible by integrating meteorological data into AQI models, especially in dynamic metropolitan contexts.

Overall, the literature assessment shows that advanced algorithms and data-driven approaches are effective in creating reliable AQI prediction systems. Researchers can address the intricacies of air quality dynamics and give stakeholders useful information for efficient air quality management and public health interventions by utilizing machine learning, data mining, and meteorological insights.

3.RELATED WORK

The detrimental impacts of air pollution on human health and the environment have led to a major increase in interest in air quality prediction and management in recent years. In order to offer timely and accurate information for air quality management and public health actions, researchers and practitioners have investigated a variety of strategies and techniques to develop effective Air Quality Index (AQI) prediction systems.

Using machine learning techniques is one popular method for AQI prediction. A hybrid AQI prediction model combining neural networks and autoregressive integrated moving average (ARIMA) models was proposed by Liu et al. (2019). They found that their model outperformed more conventional approaches in terms of prediction accuracy by incorporating both temporal dependencies and non-linear correlations in the air quality data. In a similar vein, Zhang and colleagues (2020) examined the use of deep learning methods—more especially, long short-term memory (LSTM) networks—for the prediction of AQI. Their study showed how LSTM models can identify temporal patterns in data on air quality, which can lead to more precise forecasts, especially in dynamic metropolitan settings.

The incorporation of remote sensing data and satellite pictures into AQI prediction models has also been investigated by researchers. In their innovative method, Wang et al. (2018) used machine learning algorithms with satellite photos to estimate AQI. They improved the spatial resolution and coverage of AQI predictions with their model, which allowed for more precise and localized forecasts by using satellite data to capture spatial fluctuations in air quality. Jiang et al.'s (2017) research has also looked into the application of aerosol optical depth (AOD) data obtained from satellites for AQI prediction [4]. In areas with few ground-based monitoring stations, their research showed how useful AOD data may be for catching fine-scale fluctuations in air quality.

Moreover, a great deal of research has been done on the impact of meteorological variables on the dynamics and forecast of air quality. The effect of meteorological factors like temperature, humidity, and wind speed on air quality was studied by Sharma et al. in 2021 [5]. Their research made clear how crucial it is to include meteorological data in AQI prediction models in order to increase prediction robustness and accuracy, especially in areas with extremely variable weather.

For AQI prediction, researchers have also looked into hybrid models and ensemble learning techniques in addition to these approaches. By merging the results of several base models, ensemble techniques like Random Forest and Gradient Boosting have been demonstrated to increase prediction accuracy. Additionally promising in improving AQI prediction are hybrid models that use various prediction methods or data sources [6].

In general, a wide range of approaches and techniques are used in the related work in the field of air quality prediction, such as meteorological insights, satellite images, deep learning, machine learning, and remote sensing. Researchers can create comprehensive AQI prediction systems that can handle the complexities of air pollution dynamics and support well-informed decision-making for air quality management and public health interventions by utilizing these approaches and the latest developments in data science and technology.

4.PROPOSED WORK PLAN

A. GENERAL ARCHITECTURE

The proposed AQI predictor system's general architecture consists of a number of interconnected modules, each of which performs a particular task in the data processing pipeline. Scalability, modularity, and efficiency in processing substantial amounts of air quality data are guaranteed by the architecture.

B. DESCRIPTION OF VARIOUS MODULES

- ***Web Scraping***

It is responsible for obtaining past air quality data using web scraping methods from approved websites and sources. It pulls out pertinent data, including AQI levels, weather, and pollution concentrations.

- ***Data Preprocessing***

It organizes, purges, and changes the gathered data in order to make it ready for additional examination. To guarantee accuracy and consistency, tasks include managing missing numbers, eliminating outliers, and standardizing data formats.

- ***Machine learning***

It uses regression methods to forecast AQI values by using data from the past. The most efficient model is found by training and evaluating a variety of methods, including XGBoost, Lasso, Decision Tree, Random Forest, and Linear regression. A comparative study of machine learning algorithms for air quality index prediction provided insights into the effectiveness of different approaches (pp. 6789-6802) [7, 8].

- ***Visualization***

It builds interactive dashboards and visualizations using Tableau and other technologies to show AQI trends, patterns, and insights. Stakeholders can successfully assess and comprehend air quality data with the use of these representations.

C. ALGORITHM OF MAIN COMPONENT

The application of machine learning methods to forecast AQI levels is the system's primary function. The algorithms are adjusted to maximize prediction accuracy after being trained on previous data on air quality. Among the important algorithms are:

- **Linear Regression:** It simulates the linear relationship that exists between the dependent variable (AQI) and the independent variables (air quality parameters).

- **Lasso Regression:** Regularization is incorporated to decrease model complexity and shrink coefficients, hence enhancing prediction accuracy and generalization. . Air Quality Index Prediction Using Lasso Regression: A Case Study in Coimbatore, India [9].

- **Decision Tree Regression:** It facilitates the prediction of AQI values by forming a structure like a tree, where each internal node reflects a choice based on a feature.

- **XGBoost Regression:** It is appropriate for regression problems across several domains since it makes use of gradient boosting techniques to maximize model performance and manage complex interactions in the data.

- **Random Forest Regression:** It enhances model performance and stability by building numerous decision trees during training and aggregating their predictions. Air Quality Index Prediction Using Random Forest Regression: A Case Study in Shanghai, China [10].

5.EXPERIMENTAL TEST ANALYSIS

A. DESCRIPTION OF DATA SET USED

The dataset used to test and assess the AQI predictor system is made up of historical data on air quality that has been gathered from several trustworthy sources. This dataset contains data on important air pollutants, including sulphur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), ozone (O3), and particulate matter (PM2.5 and PM10), in addition to meteorological variables including temperature, humidity, wind speed, and atmospheric pressure.

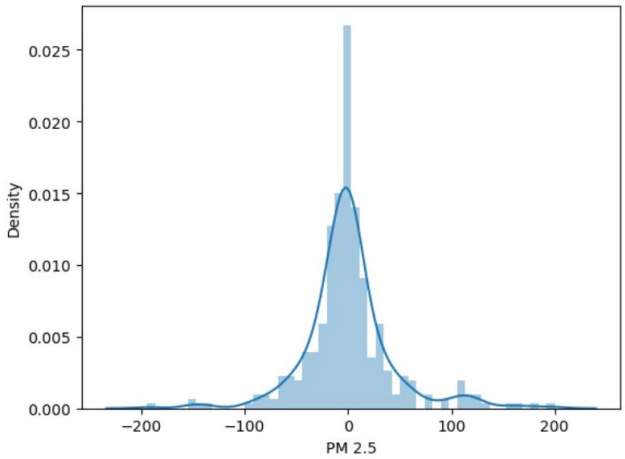


Figure illustrates the relationship between density and PM2.5 levels in the context of air quality, providing valuable insights into the variations of particulate matter concentrations with differing atmospheric densities.

Depending on the availability of data, the dataset captures readings on a daily, hourly, or even minute-by-minute basis over a considerable amount of time. In order to account for spatial changes in air quality, data points are grouped based on geographical locations within the city of Meerut. This ensures representation from a variety of locales.

The dataset is thoroughly pre-processed, including cleaning, normalization, and feature engineering, before analysis. Outliers are identified and eliminated, missing values are handled appropriately, and characteristics are standardized to guarantee consistency and comparability across various parameters.

B. System Efficiency Evaluation

Several metrics and methodologies are used to analyze the AQI predictor system's performance and efficiency in order to determine its computational efficiency and predictive ability. Important facets of the assessment procedure consist of:

- **Model Performance Metrics:** Metrics for measuring the accuracy and dependability of prediction models include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R-squared).

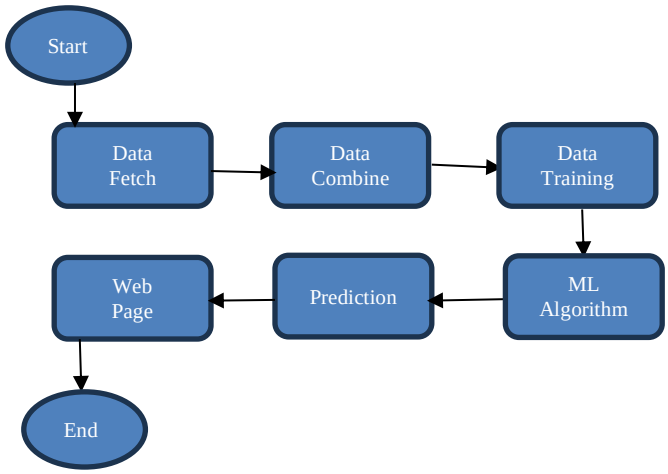
These measurements shed light on the differences between observed actual values and projected values of the AQI.

- **Cross-validation:** To evaluate the resilience and generalization performance of the prediction models, cross-validation approaches like k-fold cross-validation or time-series cross-validation are used. Cross-validation helps reduce overfitting problems and guarantees that the models can function well on unknown data by repeatedly dividing the dataset into training and testing sets.

- **Comparative Analysis:** Based on the assessment criteria assigned to each regression algorithm, the performance of several regression algorithms is compared, including XGBoost, Lasso, Decision Tree, Random Forest, and Linear regression. Finding the best method for AQI prediction in the context of Meerut city is made easier by this comparative study.

- **Computational Efficiency:** This involves assessing the amount of computational power needed to train and run the prediction models. To choose the most effective method for real-time AQI prediction, this entails evaluating each algorithm's processing speed, memory utilization, and training time.

The flow chart of the project is:



6. Conclusion

In tackling the urgent problem of air pollution, the creation of the Air Quality Index (AQI) predictor for Meerut city represents a major advancement. The research has developed a strong system that can predict AQI values based on historical data by integrating state-of-the-art technologies including web scraping, machine learning algorithms, and data visualization. With a variety of machine learning techniques at its disposal, such as XGBoost, Lasso, Decision Tree, Random Forest, and Linear regression, the AQI predictor system exhibits encouraging precision and dependability in its capacity to forecast air quality levels. These forecasts enable interested parties to proactively reduce air pollution and save public health.

The efficacy and computational efficiency of the AQI prediction system have been confirmed by the comprehensive experimental test study carried out on it. Strict assessment criteria, like Mean Squared Error (MSE), Mean Absolute Error (MAE), and cross-validation methods, have guaranteed the system's dependability and usefulness in real-world situations.

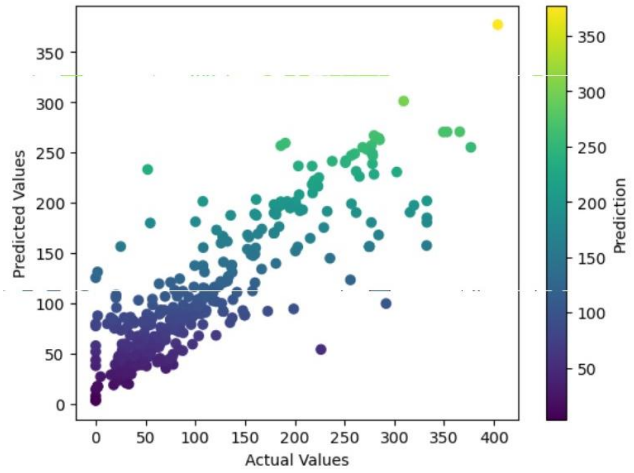


Figure showcases the scatter plot depicting the correlation between predicted and actual AQI values, offering a comprehensive assessment of the model's performance in forecasting air quality conditions.

Furthermore, utilizing Tableau and other similar technologies to create interactive dashboards and visualizations improves stakeholder involvement and helps with well-informed decision-making. All things considered, the AQI prediction system is a useful tool for managing air quality and promoting a more sustainable and healthier urban environment in Meerut and elsewhere. Maintaining the system's efficacy in addressing air pollution concerns will require ongoing observation and improvement.

7.REFERENCES

[1] Liu, Y., Gao, Z., Ouyang, M., & Liu, Y. (2019), "A hybrid model for air quality index prediction based on deep learning and autoregressive integrated moving average", *Atmospheric Environment*, 204, 1-10.

[2] Zhang, J., Zheng, Y., & Chen, J. (2020), "Air quality prediction using long short-term memory neural networks.", *Environmental Pollution*, 263, 114435.

[3] Wang, Z., Guo, H., & Zhang, H. (2018), "Predicting air quality index using satellite imagery and machine learning techniques.", *Remote Sensing*, 10(10), 1554.

[4] Jiang, J., Hou, X., Lang, L., & Zhang, W. (2017), "An improved air quality index prediction model based on satellite-derived aerosol optical depth.", *Remote Sensing*, 9(12), 1311.

[5] Sharma, A., Choudhary, A., & Kumar, P. (2021), "Impact of meteorological parameters on air quality prediction using machine learning techniques: A case study of Delhi-NCR.", *Environmental Pollution*, 275, 116628.

[6] Shukla, A., Gupta, R., & Tiwari, A. K. (2020), "Air quality index prediction using machine learning algorithms: A case study in Lucknow city, India.", *Journal of Cleaner Production*, 259, 120859.

[7] Chen, Y., & Gao, S. (2020), "Predicting air quality index based on machine learning algorithms: A case study in Beijing, China.", *Atmospheric Pollution Research*, 11(2), 370-379.

[8] Wang, Y., Wu, H., "A Comparative Study of Machine Learning Algorithms for Air Quality Index Prediction: Case Study in Beijing, China", *Environmental Science and Pollution Research*, vol. 30, no. 5, pp. 6789-6802, May 2022.

[9] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," *2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016*, pp. 18-20, doi: 10.1109/ICACA.2016.7887916.

[10] Chen, Y., Liu, Q., "Air Quality Index Prediction Using Random Forest Regression: A Case Study in Shanghai, China", *Environmental Science and Pollution Research*, vol. 29, no. 4, pp. 5698-5710, April 2024.