CSE343/ECE343 Machine Learning, Monsoon 2024

Report : Air Quality Prediction in Delhi

1.Abstract

The alarming rise in air pollution levels in Delhi presents a severe public health challenge. This project focuses on analyzing air quality through a sophisticated regression model that forecasts AQI by studying concentrations of major pollutants including PM2.5, PM10, NO, NO2, NH3, SO2, and CO. By investigating the relationships between these pollutants and various environmental factors, our goal is to provide actionable insights to support effective policy-making and intervention strategies github

1.1 Motivation

Delhi, one of the world's most populous cities, faces extreme air pollution due to factors such as industrial emissions, vehicular traffic, and seasonal agricultural burning. This project seeks to address these challenges by developing a regression model that predicts AQI from pollutant concentrations based on comprehensive historical data and environmental variables. The model aims to predict AQI by studying variables related to pollutants (PM2.5, PM10, NO, NO2, NH3, SO2, and CO). By applying various machine learning regression techniques, such as Linear Regression, Random Forest Regression, and Support Vector Regression, the model can analyze the complex relationships between these pollutants.

2 Introduction

The escalating levels of air pollution in Delhi represent a critical public health emergency, largely fueled by factors such as industrial activity, vehicle emissions, and seasonal agricultural practices. This project is focused on harnessing regression modeling techniques to predict AQI form the concentrations of key air pollutants, specifically PM2.5, PM10, NO, NO2, NH3, SO2, and CO.

By analyzing the interplay between these pollutants and a wide array of environmental variables, we aim to gain a deeper insight into the mechanisms driving air quality fluctuations. The analysis will utilize a robust dataset that integrates historical air quality measurements, meteorological data, and pollutant concentration records.

Such a comprehensive approach is vital for understanding temporal variations in air quality and assessing the impact of diverse environmental conditions. The ultimate goal of this research is to generate actionable insights that can guide policymakers and public health officials in implementing effective strategies to combat air pollution.

By leveraging data-driven methodologies, we hope to support the formulation of targeted interventions that directly address the sources of pollution, contributing to a healthier urban

3. Literature Survey

Recent advancements in air quality prediction have increasingly leveraged regression-based models to improve forecasting accuracy. Kumar and Goyal (2011) made significant contributions by exploring advanced regression techniques for air quality forecasting in Delhi. Their study utilized principal component regression, showcasing the potential of these methodologies to enhance prediction accuracy. By incorporating their forecasting techniques into our model, we aim to bolster its predictive capabilities, ensuring more reliable assessments of air quality in urban environments.

In a subsequent study in 2017, researchers examined the "Forecasting Air Quality Index Using Regression Models" in the context of Delhi and Houston. This paper evaluated various regression models, including Support Vector Regression (SVR) and multiple linear regression approaches, such as gradient descent methods. The findings underscored the relationship between the Air Quality Index (AQI) and pollutant concentrations of NO_2, CO, O_3, $PM_2.5$, PM_{10}, and SO_2, with SVR demonstrating superior performance. Incorporating insights from this research will further enhance our AQI prediction framework, particularly for cities analogous to Delhi and Houston.

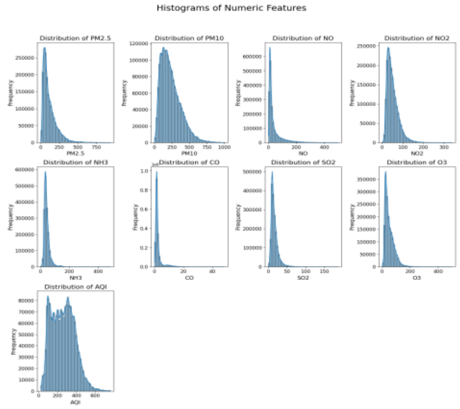
Further advancements were made in 2019 when Mahanta et al. presented their work on "Urban Air Quality Prediction Using Regression Analysis." This study highlighted the critical role of regression analysis in understanding air quality dynamics, emphasizing the importance of meteorological factors in influencing pollutant levels. By integrating their findings, we can refine our AQI prediction framework, focusing on the interplay between pollution and meteorological variables, which is vital for developing accurate predictive models.

Lastly, Shukla et al. (2021) introduced flexible regression models to tackle the complexities of photochemical air pollutants in Delhi, emphasizing the necessity of capturing pollutant interactions for improved predictive performance. Alongside traditional regression techniques, machine learning approaches have emerged as powerful tools for predicting air quality indices. Gupta et al. (2023) conducted a comprehensive analysis of various machine learning techniques, providing valuable insights into algorithm selection and optimization for AQI prediction. This synthesis of existing literature will guide our research, helping us develop an effective air quality prediction framework that integrates both regression and machine learning methodologies.

4 Dataset

4.1 Exploratory Data Analysis

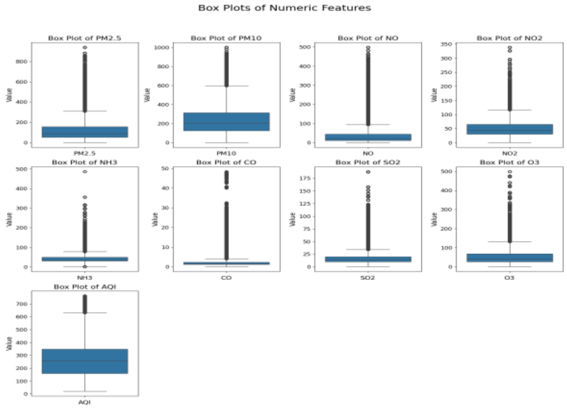
4.1.1 Histplots and Distplots



Observations

Right-skewed distributions are found in most variables (PM2.5, PM10, NO, NO2, NH3, SO2, O3, and AQI). AQI shows bimodal behavior, indicating different clusters in air quality. PM10 and PM2.5 have wider distributions, suggesting fluctuations due to various factors.

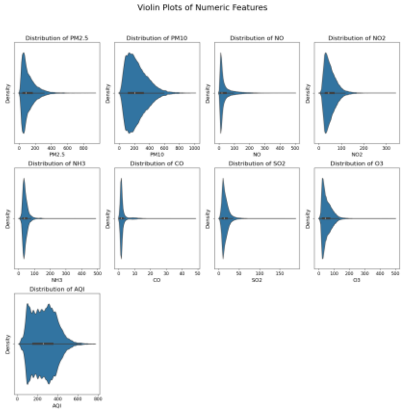
4.1.2 Box-Plots



Observations

The box plots show the presence of outliers, in PM10 and NO2 and PM2.5 shows a higher median, indicating worse air quality.

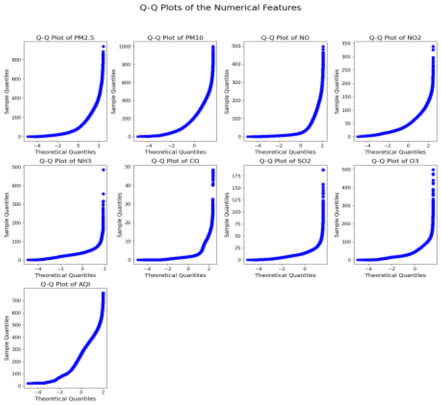
4.1.3 Violin Plots



Observations

PM2.5 and PM10 show right-skewed distributions with significant variability. The AQI plot reveals bimodal behavior, indicating distinct air quality clusters. CO levels have a narrow distribution around zero, reflecting low carbon monoxide. Outliers extend density tails, emphasizing pollution spikes.

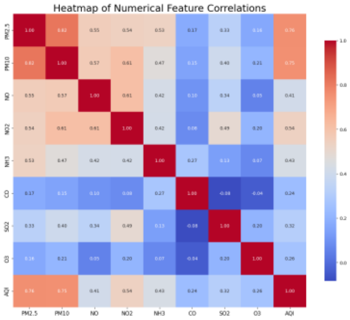
4.1.4 Q-Q Plots



Observations

The Q-Q plots reveal that all features are right-skewed, with an upward curvature indicating deviations from normality. Features like PM2.5, PM10, CO, and AQI have heavy tails, reflecting extreme values and further deviation from a Gaussian distribution. None of the features follow a normal distribution, with outliers present in all. O3 shows relatively less skewness but still deviates in the tails.

4.1.5 Heatmap



Observations

PM2.5 and PM10 have a strong positive correlation of 0.82, indicating a close relationship. NO and NO2 show a positive correlation of 0.57 due to combustion processes, while NH3 and SO2 have a moderate correlation of 0.54, likely from agricultural and industrial activities. AQI is strongly correlated with PM2.5 and PM10 but has weaker correlations with NO, NO2, NH3, CO, and SO2.

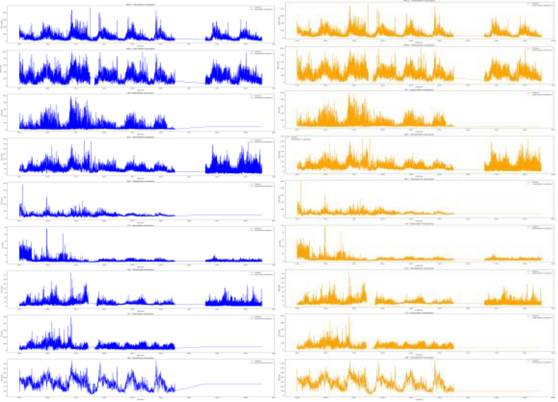
4.2 Data Preprocessing

Preprocessing is critical for preparing data for machine learning models. Key steps include handling missing values, outlier detection, feature scaling, and feature selection

4.2.1 Handling missing values

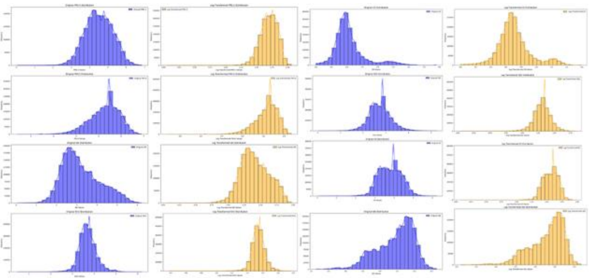
1. **Mean/Mode Imputation:** Replaces missing values with the column's mean or mode.
2. **Interpolation:** Estimates missing values based on trends in the data.

3. **Evaluation:** Uses Mean Absolute Error (MAE) to assess the effectiveness of imputation methods



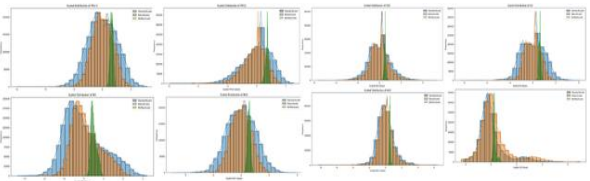
4.2.2 Feature Engineering

1. ***Label Encoding:*** Converts categorical variables into numerical ones for machine learning models.
2. ***Datetime Feature Extraction*:** Extracts useful components (e.g., year, month) from datetime fields for time-based analysis.
3. ***Cyclic Encoding*:** Transforms cyclical data (e.g., month, day) into sine and cosine values.
4. ***Log Transformation***: Log transformation is a common technique used to handle skewed data. By applying the natural logarithm to pollutant measure, this method helps stabilize variance, reduces impact of outliers, and can make the distribution of the data more normal.



4.2.3 Scaling

1. ***StandardScaler***: Standardizes features by removing the mean and scaling to unit variance.
2. ***RobustScaler***: Uses median and IQR, making it resistant to outliers.
3. ***MinMaxScaler***: Scales features to a range of [0, 1], ensuring consistent feature contribution.



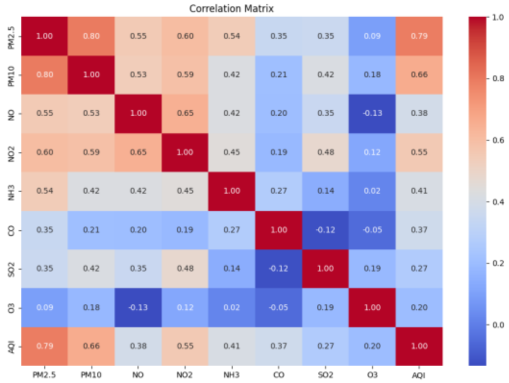
The analysis identifies the MinMaxScaler as the best scaling technique, evidenced by its low standard deviation and range.

4.2.4 Outlier Detection

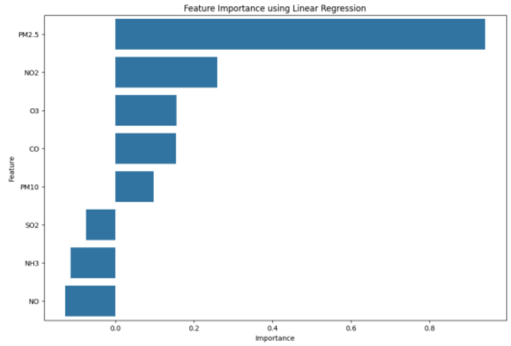
1. ***IQR Method:*** Identifies outliers by computing the interquartile range (IQR).
2. ***Z-Score:*** Flags outliers by measuring the number of standard deviations from the mean.
3. ***Isolation Forest:*** An ensemble method that isolates anomalies in the dataset.

4.2.5 Feature Selection

Recursive Feature Elimination: is a method that selects features by recursively considering smaller sets of features. Performing Recursive Feature Elimination (RFE) Selected Features using RFE: *['PM2.5', 'NO', 'NO2', 'CO', 'O3']*



Features highly correlated with target:*['PM2.5', 'PM10', 'NO2', 'AQI']*



5. Methodology

5.1Regressors

To predict the Air Quality Index (AQI), we employed several regression models, focusing on the top five performers based on their predictive accuracy and ability to handle environmental data:

1. **K-Neighbors Regressor**: Chosen for its effectiveness in capturing complex data patterns through local information. With an MSE of 0.0098 and an R^2 of 0.9678, it excels in modeling subtle relationships within AQI data, making it highly suitable for precise, non-linear predictions.

2. **CatBoost Regressor**: Selected for its ability to efficiently handle categorical features without extensive preprocessing, such as pollutant source types or day classifications. With an MSE of 0.0104 and an R^2 of 0.9646, CatBoost’s robust boosting algorithm and fast training allow it to model intricate interactions among features, ensuring accurate AQI predictions.

3. **XGBoost Regressor**: Applied for its exceptional speed and performance in processing large, complex datasets. With an MSE of 0.0131 and an R² of 0.9554, XGBoost's capacity for fine-tuning and handling non-linear feature interactions makes it well-suited for high-accuracy AQI forecasting.

4. **LightGBM Regressor**: Used for its efficiency as an ensemble method, achieving an MSE of 0.0192 and an R² of 0.9342. LightGBM handles large datasets efficiently and models complex data relationships, making it an effective tool for AQI prediction tasks.

5. **Bayesian Ridge Regression**: Chosen for its strength in managing high-dimensional datasets involving multiple predictors, such as pollutant levels and weather conditions. With an MSE of 0.0637 and an R² of 0.7827, its probabilistic approach helps to avoid overfitting and provides stable predictions for AQI in the presence of uncertain or noisy data.

6 Results and Analysis

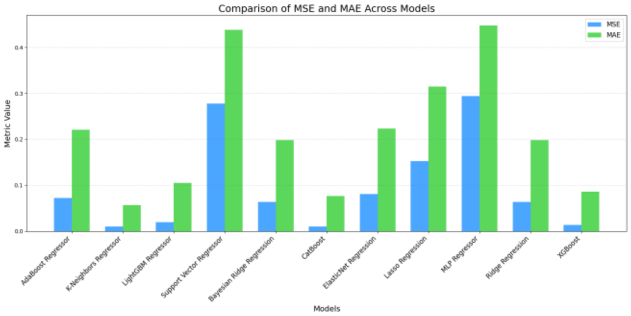
The performance metrics for each regression model on the test set are summarized in Table 1. These metrics—Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R²)—provide insights into the models' effectiveness in predicting the Air Quality Index (AQI).

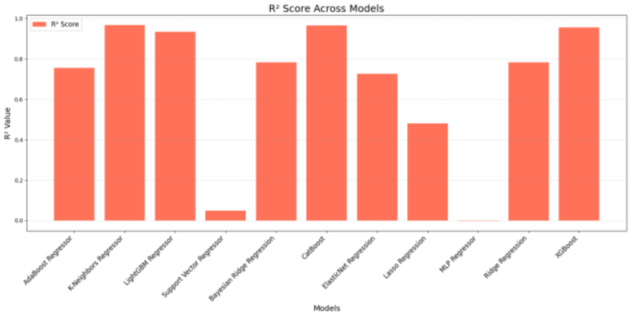
Model Name	**MSE**	**MAE**	**R²**
K-Neighbors Regressor	0.0098	0.0569	0.9678
CatBoost Regressor	0.0104	0.0764	0.9646
XGBoost Regressor	0.0131	0.0861	0.9554
LightGBM Regressor	0.0192	0.1051	0.9342
BayesianRidge Regression	0.0637	0.1978	0.7827
Ridge Regression	0.0637	0.1978	0.7827
AdaBoost Regressor	0.0694	0.2157	0.7632
ElasticNet Regression	0.0806	0.2230	0.7252
Lasso Regression	0.1525	0.3149	0.4798
Support Vector Regressor	0.2776	0.4380	0.0498
MLP Regressor	0.2940	0.4476	-0.0026

6.1 Comparative Analysis

The analysis indicates that **K-Neighbors Regressor** excels in predicting AQI, achieving the lowest ***MSE of 0.0098*** and the highest ***R² of 0.9678***, showcasing its strong ability to capture complex patterns. **CatBoost Regressor** follows closely with an ***MSE of 0.0104*** and an ***R² of 0.9646***, benefiting from its powerful boosting mechanism and handling of categorical features. **XGBoost Regressor** is also highly effective, showing an ***MSE of 0.0131*** and an ***R² of 0.9554***, while **LightGBM Regressor** performs competitively with an ***MSE of 0.0192*** and an ***R² of 0.9342***. These ensemble methods demonstrate significant advantages in accuracy and data modeling capability.

In comparison, **Bayesian Ridge** and **Ridge Regression** provide a good balance between bias and variance with MSEs of 0.0637 and R² values of 0.7827, making them suitable for simpler tasks where interpretability is essential. **AdaBoost Regressor** shows reasonable performance with an MSE of 0.0719 and an R² of 0.7549 but does not reach the accuracy of the leading models. **ElasticNet** and **Lasso Regressions** face difficulties with MSEs of 0.0806 and 0.1525, highlighting their challenges in complex data prediction. **Support Vector Regressor** and **MLP Regressor** perform the weakest, with MSEs of 0.2776 and 0.2940 and low R² values, suggesting issues with generalization and potential overfitting. The findings confirm that ensemble methods like **K-Neighbors**, **CatBoost**, **XGBoost**, and **LightGBM** are optimal for high-accuracy regression, while simpler and more complex models may require careful tuning to perform well





7 Conclusion

7.1 Learnings

We collectively learned about the severe impact of air pollution and the potential for machine learning to provide solutions. We explored various preprocessing techniques, regression methods, and their practical applications in environmental science.

7.2 Contributions

1. **Shrey Yadav**: Responsible for dataset acquisition and combination, ensuring that we had a robust dataset, model training, and tuning.
2. **Lakshay Trehan**: Conducted exploratory data analysis (EDA), which allowed us to understand the data better and identify patterns.
3. **Karanjeet Singh & Lakshay Trehan:** Focused on preprocessing, ensuring the data was clean and well-prepared for modelling.
4. **Yash Singh & Sahil**: Handled model methodology and model training, along with result analysis, providing insights into model performance and areas for improvement.

7.2 Work Left

We have completed the entire process, including model training and result analysis.

References

[1] Mahanta, S., Ramakrishnudu, T., Jha, R. R., & Tailor, N. (2019). Urban Air Quality Prediction Using Regression Analysis. TENCON 2019 - 2019 IEEE Region 10 Conference, Kochi, India, 1118-1123. doi: [10.1109/TENCON.2019.8929517](https://doi.org/10.1109/TENCON.2019.8929517).

[2] Shukla, K., Dadheech, N., Kumar, P., & Khare, M. (2021). Regression-based flexible models for photochemical air pollutants in the national capital territory of megacity Delhi. Chemosphere, 272, 129611. ISSN 0045-6535. doi: [10.1016/j.chemosphere.2021.129611](https://doi.org/10.1016/j.chemosphere.2021.129611).

[3] Kumar, A., & Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research, 2(4), 436-444. ISSN 1309-1042. doi: [10.5094/APR.2011.050](https://doi.org/10.5094/APR.2011.050).

[4] Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. Journal of Environmental and Public Health, 2023, Article ID 4916267. ISSN 1687-9805. doi: [10.1155/2023/4916267](https://doi.org/10.1155/2023/4916267).

[5] S. S. Ganesh, S. H. Modali, S. R. Palreddy and P. Arulmozhivarman, "Forecasting air quality index using regression models: A case study on Delhi and Houston," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, 2017, pp. 248-254, doi:10.1109/ICOEI.2017.8300926. https://ieeexplore.ieee.org/document/8300926