CSE343/ECE343 Machine Learning, Monsoon 2024
Mid-Sem Project Evaluation

Lakshay Trehan 2022267 · Karanjeet Singh 2022235 · Sahil 2022427 · Shrey Yadav 2022483 · Yash Singh 2022589

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.

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 environment for the residents of Delhi.

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.

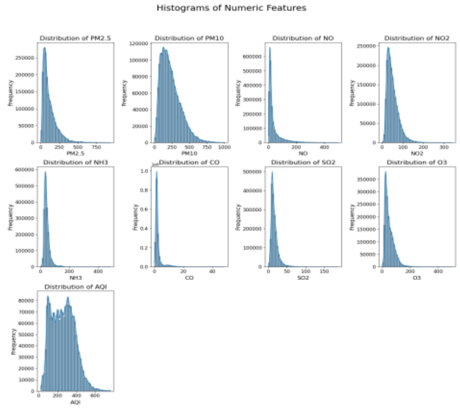
Further advancements were made in 2019 when Mahanta et al. presented their work on "Urban Air Quality Prediction Using Regression Analysis." This study highlighted the critical role of regression analysis in understanding air quality dynamics, emphasizing the importance of meteorological factors in influencing 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

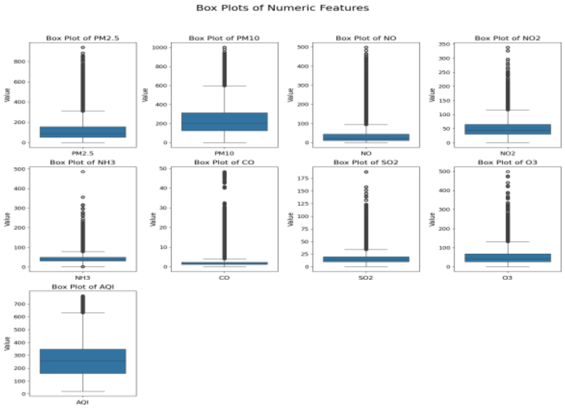
4.1.1 Histplots and Distplots



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.

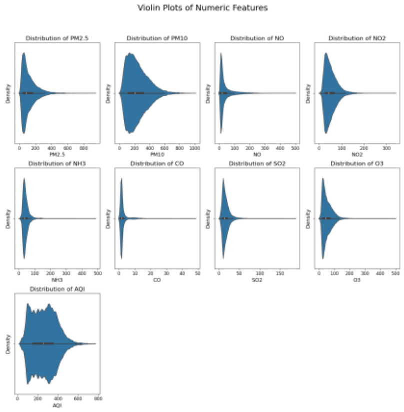
4.1.2 Box-Plots



Observations

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

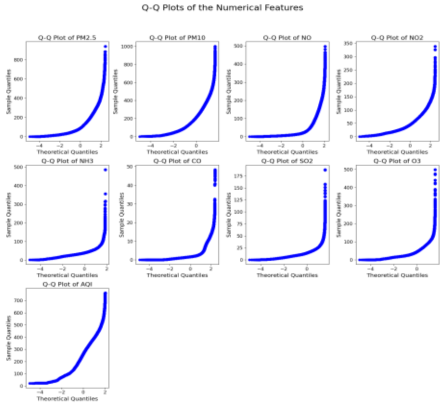
4.1.3 Violin Plots



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.

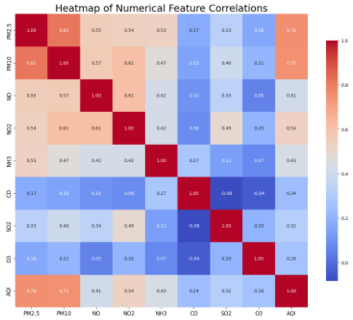
4.1.4 Q-Q Plots



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.

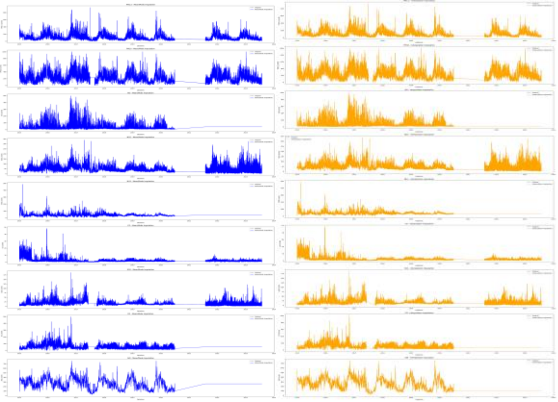
4.2 Data Preprocessing

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

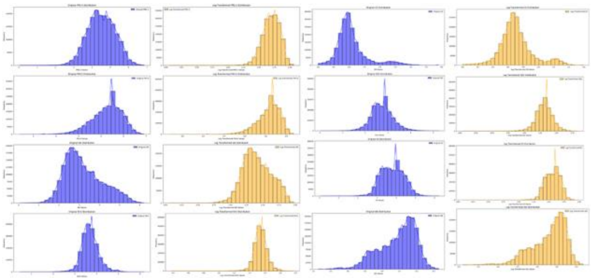
1. **Mean/Mode Imputation:** Replaces missing values with the column’s mean or mode.
2. **Interpolation:** Estimates missing values based on trends in the data.

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



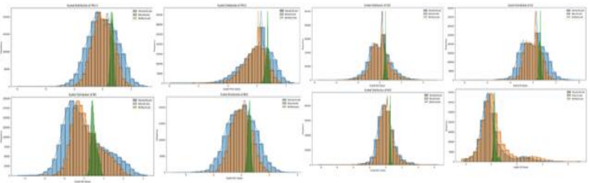
4.2.2 Feature Engineering

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.



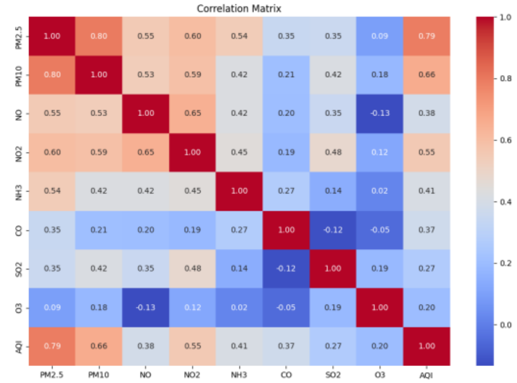
The analysis identifies the MinMaxScaler as the best scaling technique, evidenced by its low standard deviation and range.

4.2.4 Outlier Detection

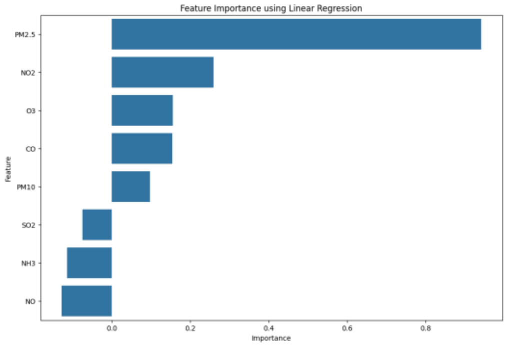
1. ***IQR Method:*** Identifies outliers by computing the interquartile range (IQR).
2. ***Z-Score:*** Flags outliers by measuring the number of standard deviations from the mean.
3. ***Isolation Forest:*** An ensemble method that isolates anomalies in the dataset.

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, each with distinct advantages for handling complex environmental data:

1. **AdaBoost Regressor**: Chosen for its capability to boost weak learners by focusing on difficult-to-predict AQI values. This makes it useful when AQI fluctuations are driven by subtle, nonlinear factors.

2. **Bayesian Ridge Regression**: Selected for its strength in managing high-dimensional AQI datasets that involve many potential predictors (e.g., pollutant levels, weather conditions). Its probabilistic approach helps avoid overfitting when dealing with uncertain or noisy AQI data.

3. **CatBoost**: Used for its efficiency in handling categorical features, such as pollutant source types or day classifications, without the need for extensive preprocessing. Its fast training

and ability to model complex relationships between features ensure accurate AQI predictions.

4. **ElasticNet Regression**: This model was employed due to its combination of L1 and L2 regularization, making it adept at handling AQI datasets with correlated features (such as pollutant concentrations that often co-occur). This regularization helps in making more robust predictions.

5. **Lasso Regression**: Used for its feature selection capabilities, allowing us to eliminate irrelevant features in AQI prediction, such as potentially redundant or low-impact variables, thus focusing on the most critical environmental factors influencing air quality.

6. **MLP Regressor (Multi-Layer Perceptron)**: Applied for its ability to capture non-linear relationships in AQI data, such as the nonlinear effect of weather conditions (temperature, humidity) on air quality. However, its performance can vary depending on hyperparameter tuning.

7. **Ridge Regression**: Chosen for its ability to handle multicollinearity among AQI-related variables (e.g., multiple pollutants). By penalizing large coefficients, it helps create more stable models, which is crucial when predicting air quality influenced by interdependent factors.

8. **XGBoost**: This model was selected for its exceptional speed and ability to handle large, complex datasets, such as those used in AQI prediction. Its capacity for fine-tuning allows it to model intricate interactions between various AQI contributors, making it ideal for high-performance AQI forecasting.

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).

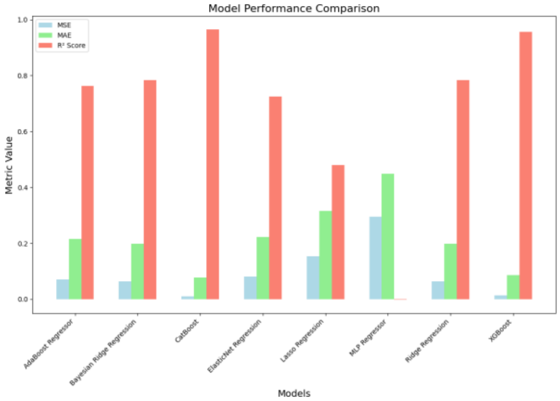
Model Name	**MSE**	**MAE**	**R²**
AdaBoost Regressor	0.0694	0.2157	0.7632
BayesianRidge Regression	0.0637	0.1978	0.7827
CatBoost Regressor	0.0104	0.0764	0.9646
ElasticNet Regression	0.0806	0.2230	0.7252
Lasso Regression	0.1525	0.3149	0.4798
MLP Regressor	0.2940	0.4476	-0.0026
Ridge Regression	0.0637	0.1978	0.7827
XGBoost Regressor	0.0131	0.0861	0.9554

6.1 Comparative Analysis

The analysis of regression models for predicting Air Quality Index (AQI) highlights the ***CatBoost Regressor*** as the top performer, achieving the lowest ***Mean Squared Error (MSE) of 0.0104*** and a ***high R² value of 0.9646***. Its effectiveness is attributed to its advanced boosting mechanism and robust handling of categorical variables, allowing it to efficiently capture intricate patterns in the AQI data.

The ***XGBoost Regressor*** follows closely with an ***MSE of 0.0131*** and an ***R² of 0.9554***, demonstrating strong capabilities in uncovering complex relationships within the data. Both ***Bayesian Ridge*** and ***Ridge Regression*** provide competitive performance, with similar MSE values around ***0.0637***, showcasing their effectiveness in balancing bias and variance, although they do not match the accuracy of the boosting methods.

In contrast, ***Lasso Regression*** and the ***MLP Regressor*** exhibited the weakest performance, with MSE values of ***0.1525*** and ***0.2940***, respectively. These models struggled with generalization and overfitting, highlighting the challenges faced by linear models and neural networks in capturing the complexity of AQI data. Overall, CatBoost's superior accuracy underscores the benefits of advanced ensemble methods in predictive modeling tasks.



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 to work with.
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

Currently, we are on track with our timeline and are now focusing on tuning ensemble models to enhance our analysis and improve predictive accuracy.

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