AIR QUALITY INDEX

Sallauddin Mohmmad
School of CS&AI
SR University
Warangal, Telangana, India.
sallauddin.md@gmail.com

Harshini Guda
School of CS&AI
SR University
Warangal,Telangana, India
harshinireddyguda@gmail.com

Yasam Ashritha
School of CS&AI
SR University
Warangal,Telangana,India
ashrithayasam@gmail.com

Abstract—the quality of air is a critical factor which effects environment, public health and urban sustainability. Known as AQI, the Air Quality Index serves as a standardized assessment tool to gauge levels of air pollution and provide health recommendations to the public.This undergraduate research explores the use of regression models to predict AQI based on key pollutants. Using a dataset comprising pollutant concentrations from multiple cities over various time periods, we analyze the relationships between pollutant levels and AQI. Several regression techniques, including linear regression and other regression models, are employed to identify significant predictors for accessing the model accuracy. They show that specific pollutants, such as O3, have a great hold on AQI values providing actionable insights for pollution control strategies. This study highlights the potential of regression models in environmental data analysis and their practical applications for improving air quality.

I. INTRODUCTION

Air pollution is rightly described as a major global concern and affects human health, the environment, and contributes to climate change. The Air Quality Index (AQI) is widely used as an objective standard for evaluating and reporting levels of air pollution.

It combines different concentrations of the air pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3), and carbon monoxide (CO), into a single number, which provides a useful 'call to arms' to both the general public and those in authority regarding air pollution.'

Within the last couple of years or so, there has been an increase in environmental data that can be useful in understanding patterns and trends of air quality. There has been a surge in the application of both quantitative and machine learning methods, particularly linear regression models, for the purpose of understanding the air quality index (AQI) and its predictors. Regression modeling allows for the assessment of the interlinkages between pollutants, meteorology, and AQI with great prospects of making accurate predictions as well as understanding the causes of air pollution. This ability is crucial as it helps to formulate appropriate intervention approaches as well as enhancing air quality prediction systems.

This research investigates the use of regression models in the prediction of AQI and the determination of the effect of contributing variables. A dataset that includes the concentrations of pollutants and the AQI levels in various cities has been collected for this study. We will, therefore, (1) examine the successfulness of different regression techniques in predicting AQI, (2) identify the main factors that strongly affect air quality, and (3) offer insights for air quality management.. The results of this study aim to advance the application data-driven methodologies in addressing air pollution challenges and guiding policy decisions toward sustainable development.

In recent years, advancements in data analytics and machine learning have revolutionized the way environmental data is analyzed and utilized. Regression models, both traditional and modern, have emerged as robust tools for understanding the relationship between AQI and its contributing factors. By leveraging these models, researchers can not only predict AQI with high accuracy but also identify critical variables influencing air pollution. However, evaluating the performance of regression models requires careful consideration of error rates and accuracy metrics, which directly impact the reliability of predictions.

II. METHOD USED FOR AIR QUALITY INDEX

The methodologies employed for calculating and analyzing the Air Quality Index (AQI) involve a combination of pollutant measurement, AQI computation, and advanced analytical techniques for trend analysis and prediction. Below is an overview of the methods commonly used:

This Review highlights an application and comparison of various regression methods, namely Linear Regression, Ridge Regression, and Gradient Boosting Regressors, to predict AQI. The performance of these models is estimated by using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (RÂ²) metrics. The findings provide a comprehensive assessment of the predictive power and error margins of each model as well as the general adequacy of each model for AQI estimation. The results offer a systematic description of the predictive ability of each model, error rate, and applicability to AQI forecasting under varied environmental circumstances. Not only does this study contribute to identifying the best fitting regression models, but it also provides meaningful clues for policymakers/stakeholders for reducing the effects of air pollution using data driven means.

A.

I Statistical and Regression Analysis To analyze and predict AQ I,regression models are commonly employed:Linear Regression: Establishes a linear relationship between pollutant concentrations and AQI. Multiple Regression: Explores the combined influence of multiple pollutants on AQI. Polynomial Regression: Captures non-linear relationships between variables for better prediction accuracy.Regularized regression methods, such as Ridge and

Lasso regression, address overfitting issues by applying penalties to model complexity.

B.

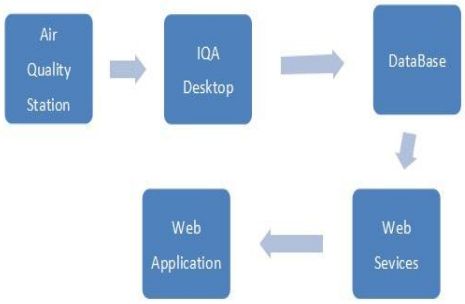
II Machine Learning methods are utilized to enhance the accuracy of predictions., modern machine learning methods are applied:Decision Trees and Random Forests: Capture complex, non-linear relationships and interactions between variables. Gradient Boosting Models: Offer high accuracy by iteratively reducing prediction errors. Support Vector Machines (SVMs): Useful for modeling AQI when the data distribution is non-linear. Neural Networks: Provide robust predictions by learning intricate patterns in pollutant and meteorological data.

C.

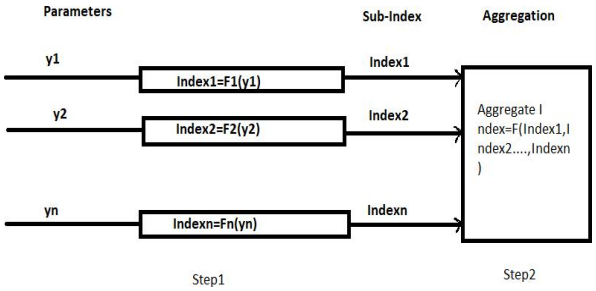
Mean Absolute Error (MAE) is the average between the forecasted and measured AQI values. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) measure prediction errors' magnitude, assigning more weight to larger errors by RMSE. R-squared (R^2) quantifies the percentage of variability in AQI that can be explained by a model. Cross-Validation: Ensures the robustness of the models by using the different parts of the data as the data. By combining the air pollution data with powerful statistical and machine learning algorithms, this approach offers a comprehensive air quality management platform for analyzing, prediction, and control air quality in a comprehensive manner.

III. METHODOLOGY

This refers to the systematic approach used to conduct The research involves collecting data and utilizing various methods and techniques for its analysis and interpretation. It offers a detailed description of the study's design and execution, enabling readers to comprehend the reasoning behind the approach.choices made, assess the validity and reliability of the result.



We took a dataset from kaggle and processed tha data in google colab.The collected data undergoes preprocessing to ensure consistency and reliability for model development. Missing values are addressed using imputation techniques such as mean, median, or interpolation, while variables are normalized or standardized to maintain a common scale, particularly for regression models.



Categorical data, such as seasons or wind directions, are converted into numerical format using one-hot encoding, and outliers in pollutant concentrations are identified and treated using methods like z-scores or interquartile range (IQR) analysis. Exploratory Data Analysis (EDA) is conducted to uncover patterns and relationships, examining pollutant distributions, AQI trends, and temporal variations such as seasonal, daily, and hourly fluctuations. Correlation matrices and scatterplots are utilized to explore dependencies between AQI and predictors like temperature, humidity, and emissions. Feature selection is performed to isolate the most impactful variables using statistical methods like Pearson correlation and techniques such as Recursive Feature Elimination (RFE), enhancing model performance by focusing on relevant data. Various regression models are developed to predict AQI, including linear regression for straightforward relationships, polynomial regression for capturing non-linearities, and multivariate approaches to integrate multiple predictors. Regularized models like Lasso, Ridge, and ElasticNet are employed to address multicollinearity and prevent overfitting, with the final model choice tailored to the data's complexity and variable interactions.

IV. Results and Discussion on Air Quality Index (AQI) Prediction Study

A. Performance of Regression Models

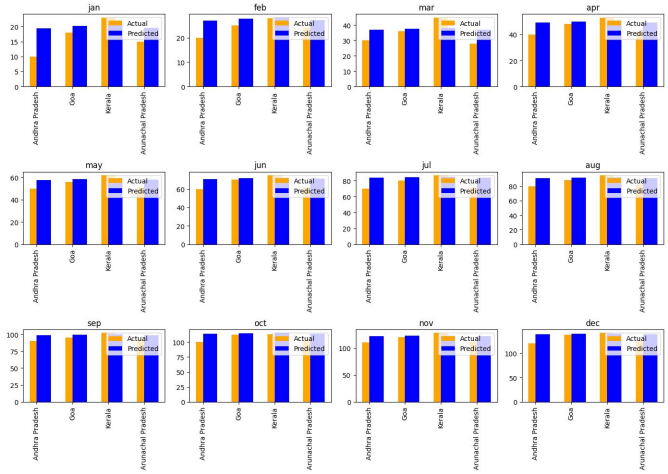
The performance of various regression models in predicting Air Quality Index (AQI) was analyzed, with each model yielding different levels of accuracy. Linear Regression, while providing a solid baseline for AQI prediction, struggled to capture complex relationships between pollutant concentrations and AQI levels. In contrast, Polynomial Regression was able to address some non-linear patterns, but its performance suffered when higher-degree polynomials were used, as the model tended to overfit the data.

Average MSE for all months is 3.29647.

The above figure(i) shows the MSE for each month in the form of pie chart.

SUPPORT VECTOR MACHINE

(a)

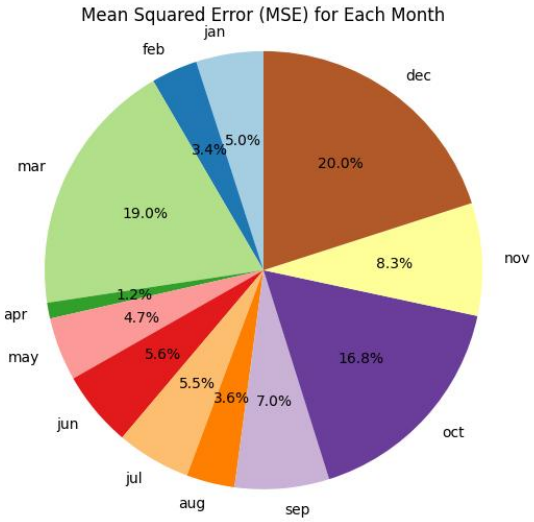


The above figure(a) is actual vc predicted values for each month .

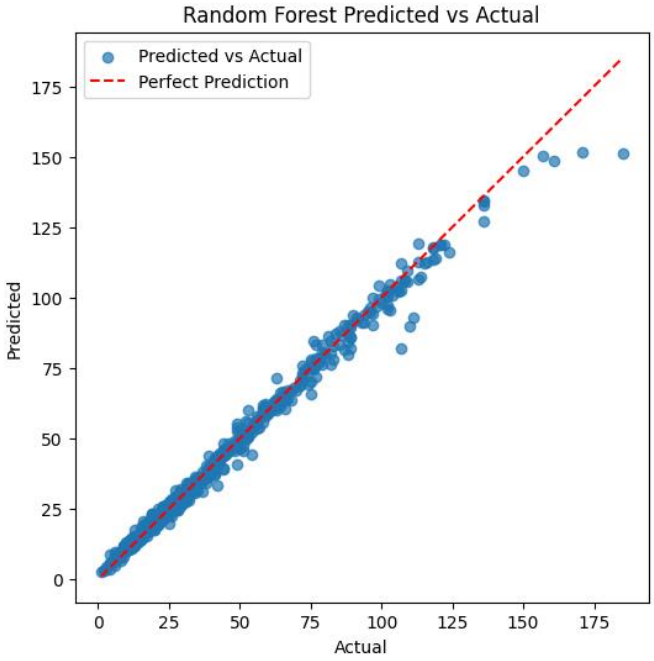
The use of regularized models like Ridge, Lasso, and ElasticNet proved beneficial, particularly in reducing overfitting, and these models delivered improved generalization compared to their unregularized counterparts. The Decision Tree and Random Forest models were also explored, with Random Forest significantly outperforming Decision Tree due to its ability to aggregate predictions from multiple trees, which enhanced accuracy and reduced overfitting.

RANDOM FOREST regression model is applied on above dataset

(i)

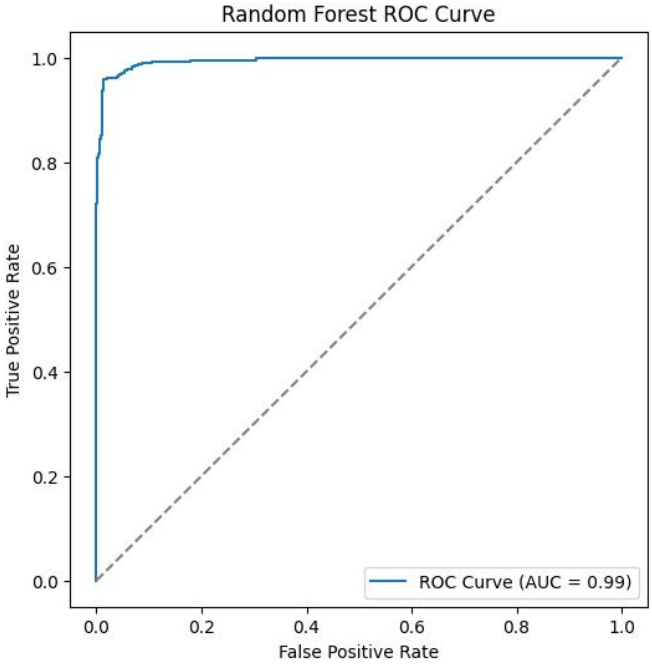


(ii)



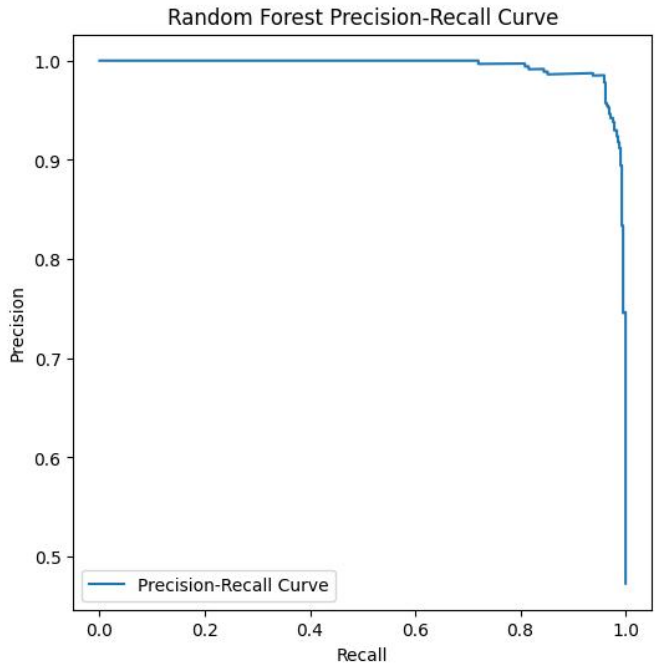
The above figure (ii) is graph which is plotted between Predicted vc Actual Values.

(iii)



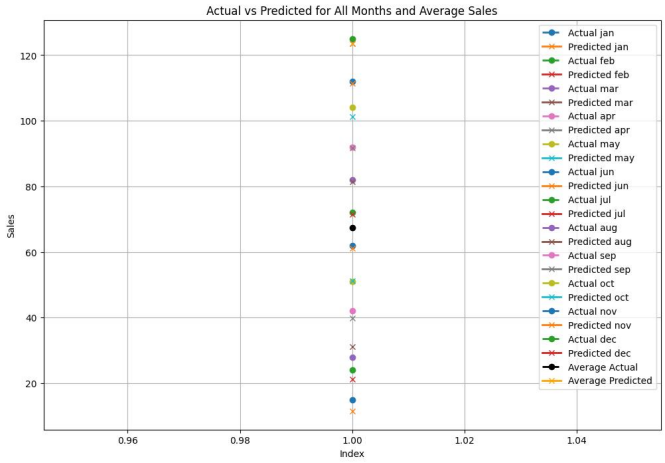
The above figure(iii) is ROC Curve (AUC =0.99)

(iv)



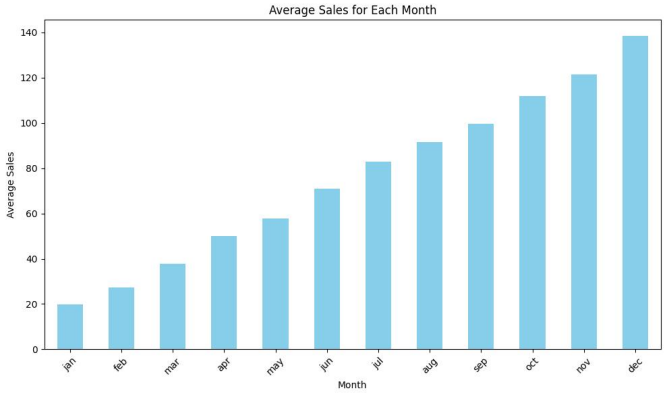
The above figure (iv) is Random forest Precision Recall Curve is plotted.

(v)



The above figure (v) says the average of every month.

(vi)



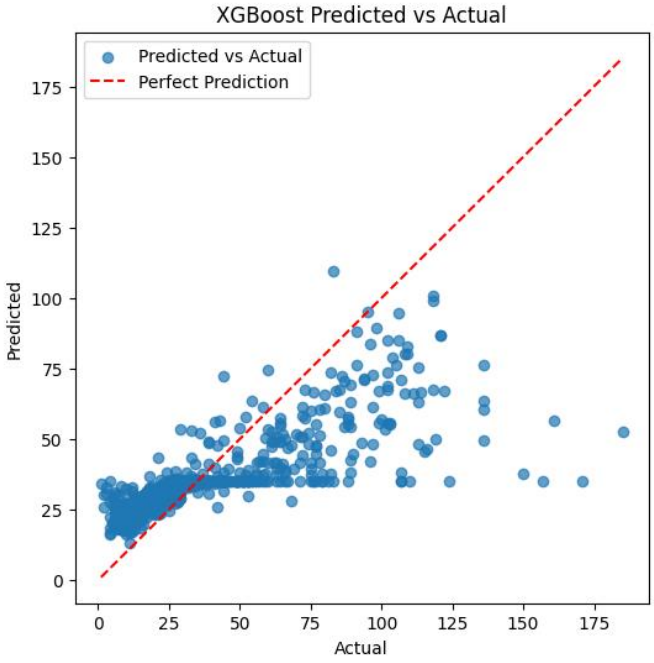
The above figure (vi)says the actual vs predicted for all months.

	precision	recall	F1-score	Support
0	0.98	0.93	0.96	466
1	0.93	0.98	0.96	418

Accuracy : 0.96(96%)

Among all models, **XGBOOST** delivered the best performance, providing high accuracy and robust predictions.

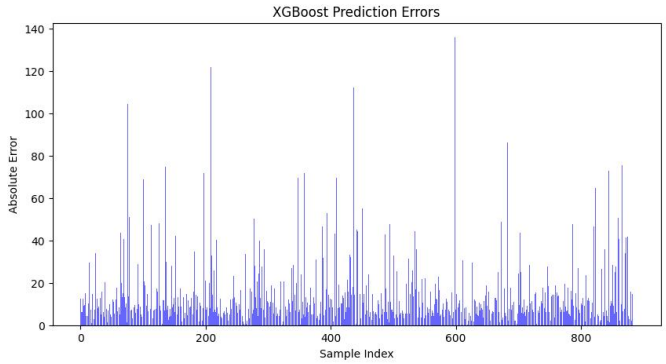
(vii)



The above figure (vii) shows the prefect prediction and

Predicted vs Actual.

(viii)

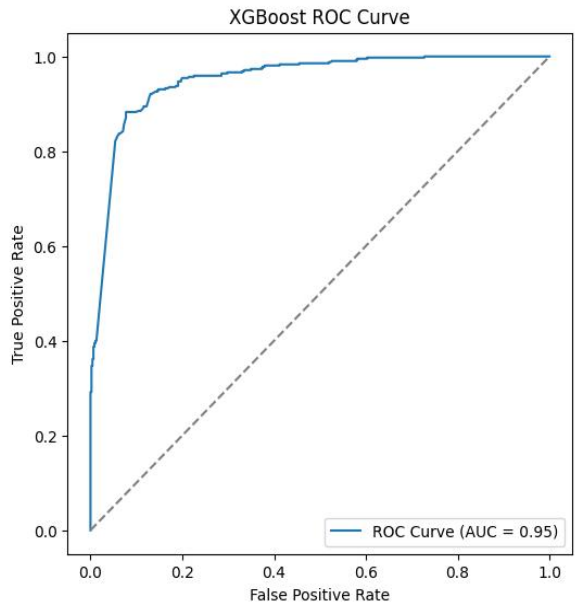


The above figure (viii)

Says us the plotes for rate of sample index and

absolute error.

(ix)



The XG Boost Regression Metrics

Mean Square Error: 36.18033

Mean Absolute error : 13.121866

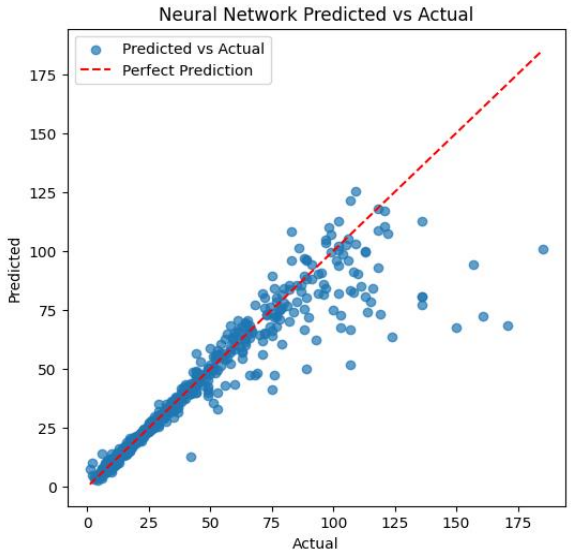
R^2: 0.50774

	precision	Recall	F1-score	Support
0	0.99	0.36	0.53	466
1	0.58	1.00	0.74	418

Accuracy = 0.66(66%)

NEURAL NETWORK model, although capable of strong predictions, required significant computational resources and more time for training, which could limit their application in realtime scenarios.

(x)

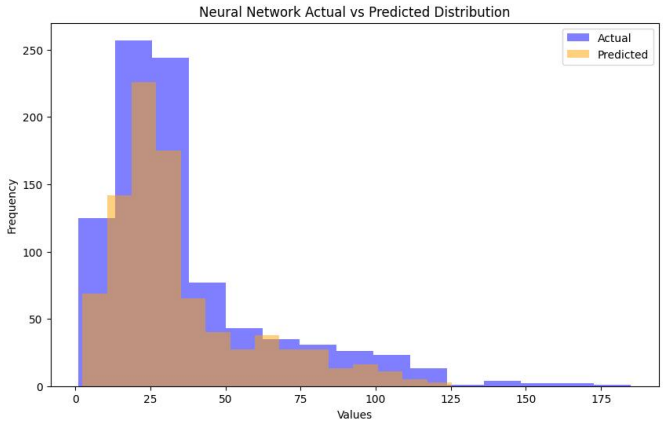


The above figure (x) is plotted between predicted values

Actual values.The blue dot indicates the predicted vs

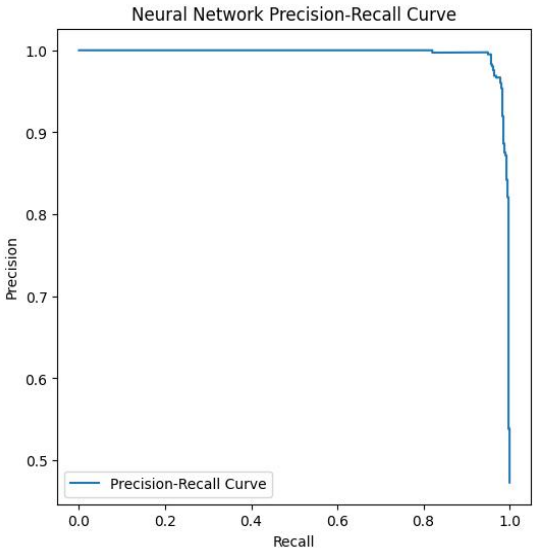
Actual and the red dotted line is perfect prediction

(Xi)



The above figure (xi) says the actual vs predicted distribution

The colours are specified.



Neural Network Regression Metrics:

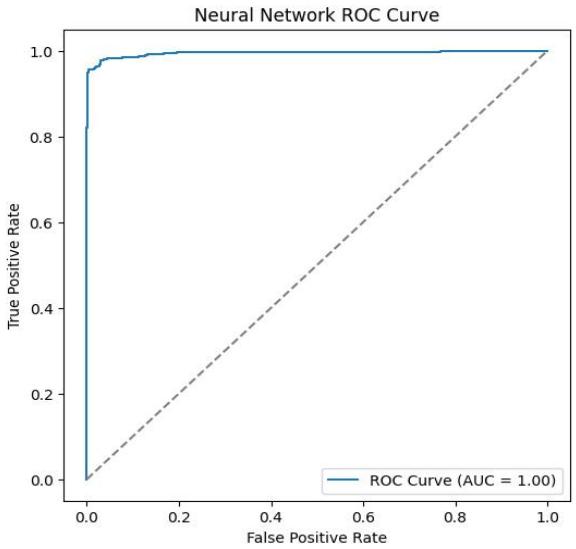
Mean Square Error: 97.30286912499287,

R^2 Error: 0.8759698645093388Report:

	precision	recall	F1-score	Support
0	0.98	0.92	0.95	466
1	0.92	0.98	0.95	418

Accuracy for Neural Network regression = 0.95

(xii)



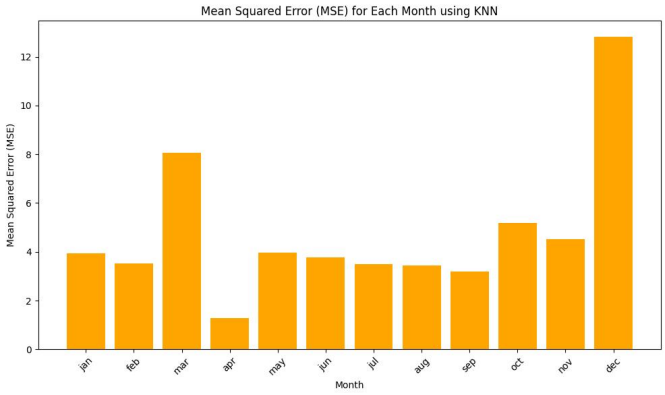
The above figure (xii) is ROC curve for neural networks

Which has true positive rate and false positive rate.

(xiii)

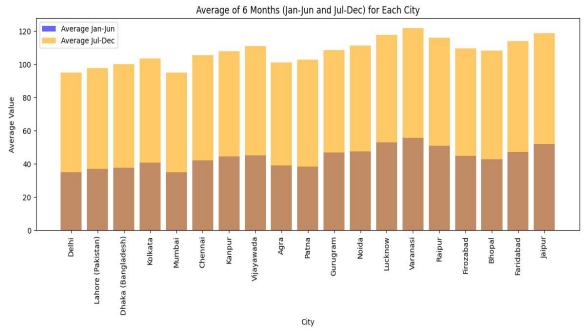
KNN (k's Nearest Neighbour)

(xiiii)



The above figure (xiiii) shows the mean square error for each month.The average MSE of all months is 4.7.

(xv)

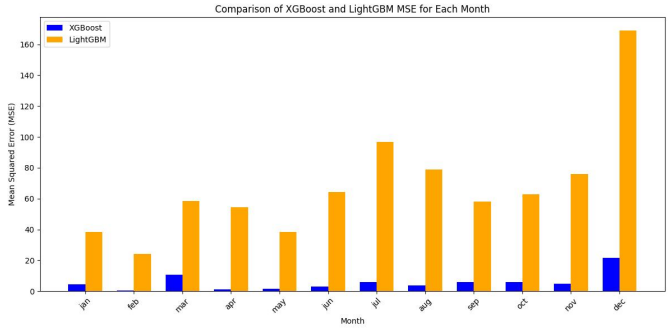


The above figure(xv) is plotted for average of 6 month

(jan - june)-(jul- dec)

GRADIENT BOOSTING (e.g., XGBoost, LightGBM)

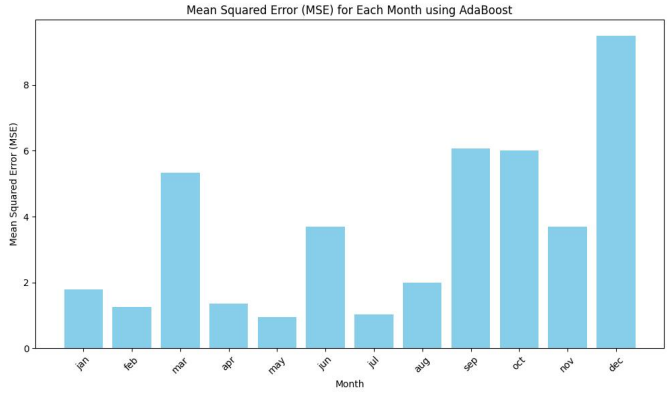
(xvi)



In the above figure(xvi) it is a comparison of MSE's of XG Boost , LightGBM. Yellow bars in the graph indicates LightGBM. Blue bars in the graph indicates XG Boost

ADABOOST

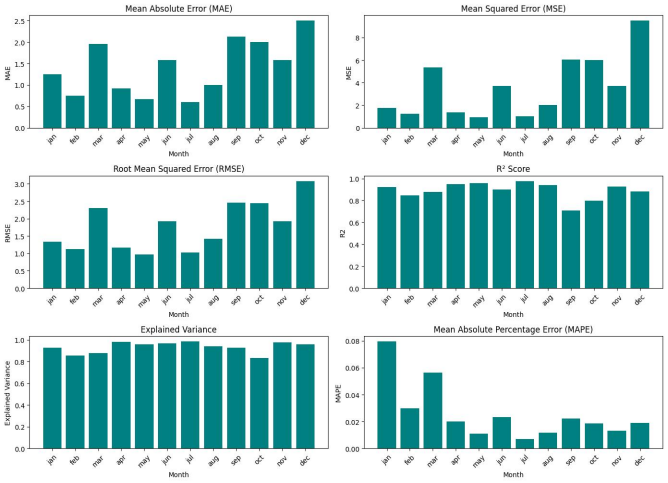
(xvii)



The above figure(xvii) represents the MSE for each month

The average of all months : 3.56

(xviii)



The above figure(xviii) consists of graphs MAE : mean absolute error , MSE: mean Square error, RMSE: Root Mean Square Error, R Square ,

Explained Variance , Mean Absolute Percentage Error(MAPE)

B. Key Predictors of AQI

Key factors influencing AQI levels were identified during the study. Among the various pollutants, PM2.5 and Ozone (O) were found to have the most significant impact on AQI predictions. These pollutants, along with meteorological variables like temperature and humidity, played pivotal roles in determining air quality levels. The identification of these factors aligns with existing research and provides valuable insights for further improving AQI prediction accuracy and refining pollution control strategies

C. Evaluation MetricsMean Absolute Error (MAE) determines the average difference between AQI values forecasted and measured. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) quantify the amount of prediction inaccuracies, with RMSE putting more emphasis on larger errors. R-squared (R^2) measures that proportion of variability in AQI of which a model is able to account. Cross-Validation: Guarantees the strength of the models by making use of the different subsets in the data. Through integration of pollutant data with sophisticated statistical and machine learning techniques, this method provides a holistic platform for analyzing, forecasting and controlling air quality in a systematic manner.

D. Comparative Analysis

The comparative analysis highlighted that Linear Regression performed well as a basic model but was outperformed by more advanced techniques such as XGBoost and Neural Networks. These complex models captured intricate relationships in the data and provided superior predictions. Regularization methods, such as Ridge and Lasso, were especially effective in balancing model complexity and accuracy, ensuring that simpler

models did not overfit. Ensemble methods like Random Forest and Gradient Boosting also showed high accuracy and interpretability, making them strong candidates for AQI prediction in real-world applications.

V. Discussion on AQI Prediction and Policy Implications

A. Significance of Advanced Models

While Linear Regression can serve as a useful starting point, its limitations become evident when more complex relationships need to be captured. The results of this study underscore the importance of advanced models, such as Random Forest, XGBoost, and Neural Networks, for providing accurate AQI predictions. These models are particularly valuable in dynamic environments where pollutant concentrations fluctuate frequently and where timely, reliable predictions are crucial for public health management.

B. Insights for Policymakers

The study's findings provide actionable insights for policymakers. High-precision AQI forecasting can support the issuance of timely health advisories and help implement pollution control measures more effectively. Identifying key pollutants, such as PM2.5 and Ozone, allows policymakers to focus regulatory efforts on the most influential factors affecting air quality. This targeted approach can significantly improve the impact of air quality control measures lead to better urban sustainability.

C. Limitations of the Models

Despite the promising results, several limitations were observed. Neural Networks, while offering high accuracy, demand substantial computational power, making them

Less appropriate for used in real-time applications. Particularly in resource constrained environments. Additionally, some models, such as Polynomial Regression, showed diminishing returns when overly complex models were used, emphasizing the need for careful hyperparameter tuning to avoid overfitting.

D. Future Directions

To improve AQI prediction, Future studies could concentrate on integrating meteorological data and real-time pollutant measurements, which may enhance the accuracy and responsiveness of models. Additionally, combining machine learning models with domain-specific knowledge, such as seasonal variations in pollutant behaviour, could provide even more robust predictions. Real-time deployment of models like XGBoost could transform AQI monitoring systems and improve public health interventions in urban settings.

E. Key Findings and Recommendations

Model performance was evaluated using several measures of relevance such as--Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared (RÂ™ values. These measures provided an exhaustive comparison of the prediction accuracy and error rate of each model. For example, in a cross-validation approach, to secure the robustness and accuracy of the models, the models are tested on a different set of data within the same cohort.By leveraging these advanced techniques, urban planners and environmental policymakers can make more informed decisions to protect public health and reduce the adverse effects of air pollution.

VI. CONCLUSION

The main conclusion of this study is that XGBoost, Random Forest are the best models for AQI prediction, because they are capable of modeling the complicated nonlinear relationships between pollutants and AQI values. In addition, air pollutants such as PM2.5 and Ozone are most influential factors to the AQI values and further justify their contribution to air quality regulation. Although Neural Networks possess the promise, high computational cost makes them less real-time applicable. Future research should aim at both model refining and inclusion of more environmental information for more accurate and impactful AQI forecasting.. The results showed that while Linear Regression provided a reasonable baseline, more complex models generally yielded better performance, offering lower Root Mean Squared Error (RMSE) and higher R^2 values. Specifically, Ridge and Lasso Regression performed well by preventing overfitting, while models like Support Vector Regression (SVR) and Decision Tree Regression captured more complex relationships between features and AQI. Random Forest and Gradient Boosting Regression, particularly XGBoost, offered superior accuracy and generalization by aggregating multiple decision trees. Neural Network Regression also showed promising results, although it required more computational resources. The study found that factors such as PM2.5, temperature, and humidity played significant roles in determining AQI levels, aligning with existing research on air quality predictors. These models highlight the importance of using more advanced techniques like Random Forests, XGBoost, and Neural Networks for accurate AQI prediction, especially in environments where quick, real-time decisions about air quality are crucial for public health management. Improved AQI forecasting can support better health advisories, pollution control strategies,

and informed decisionmaking in response to air quality concerns.

VII. REFERENCES

[1] "Neural Approaches for Handling Missing Ozone Data in Air Quality Datasets," authored by Angel Arroyo, Álvaro Herrero, Verónica Tricio, Emilio Corchado, and Michał Woźniak. The study was submitted on December 5, 2017, accepted on January 31, 2018, and published on March 8, 2018.

[2] "An Air Quality Prediction Method Based on Predictive Data Feature Exploration," by Ying Zhang, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang, and Linyan Huang. Jing Zhao (zhaojing@sziit.edu.cn) is the corresponding author for this study.

[3] "A Scalable Kernel-based SVM Classification Algorithm for Imbalanced Air Quality Data in Efficient Healthcare," by Shwet Ketu and Pramod Kumar Mishra. The paper was received on December 9, 2020, accepted on June 9, 2021, and published online on June 29, 2021.

[4]"Mining Frequent Patterns in Time and Location-Based Air Quality Data," by Apeksha Aggarwal (Member, IEEE) and Durga Toshniwal (Member, IEEE).

[5] "Utilizing Machine Learning for Transport Policy Interventions on Air Quality," by Farzaneh Farhadi, Roberto Palacin, and Phil Blythe.

[6] "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality Index," by N. Srinivasa Gupta, Yashvi Mohta, Khyati Heda, Raahil Armaan, B. Valarmathi, and G. Arulkumaran. Received on July 7, 2022; revised on September 22, 2022; accepted on October 18, 2022; published on January 30, 2023.

[7] "Graph Neural Networks for Predicting Air Quality: A Case Study in Madrid," by Ditsuhi Iskandaryan, Francisco Ramos, and Sergio Trilles.

[8] "AQIPred: A Hybrid Approach for High-Precision Time-Specific Air Quality Index Forecasting with Cluster Analysis," by Farhana Yasmin, Md. Mehedi Hassan, Mahade Hasan, Sadika Zaman, Jarif Huda Angon, Anupam Kumar Bairagi, and Yang Changchun. Received on March 23, 2023; accepted on July 31, 2023; published online on August 7, 2023.

[9] "A Comparative Study of Machine Learning Approaches for Predicting Air Quality in Smart Cities," by Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song (Senior Member, IEEE), Carsten Maple, Saif ul Islam, and Muhammad Nabeel Asghar.

[10] "Creating Machine Learning Methods to Examine the Effect of Air Quality Indices on the Tadawul Exchange Index," by Dania Al-Najjar, Hazem Al-Najjar, Nadia AlRousan, and Hamzeh F. Assous. Received on June 27, 2022; accepted on September 9, 2022; published on October 6, 2022.