iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68825

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Air Quality Index Prediction Using Ensemble Learning

Shraddha Ishwarchandra Ghonsikar[1], Pravin R. Rathod[2]
Computer Science and Engineering, Dr. Babasaheb Ambedkar Technological University

Abstract: ***In smart cities, air pollution has harmful impacts on human physical health and the quality of living environment. correctly predicting air quality is important for developing effective strategies to reduce air pollution and promote healthier, more sustainable environments. Tracking and predicting air pollution is essential for enabling individuals to make well-informed choices that safeguard their health. Predicting air quality is vital for public health, environmental management, and the development of effective policies. This research focuses on predicting the Air Quality Index (AQI) using machine learning techniques, with an emphasis on improving model efficiency and prediction accuracy. This study compares the performance of two regression algorithms: Linear Regression and Principal Component Regression (PCR) with Decision Tree Regression, across several key evaluation metrics. The performance is assessed using measures such as Mean Squared Error (MSE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Median Absolute Error (MedAE), Explained Variance, Adjusted R^2. The analysis reveals that the PCR with Decision Tree Regression model outperforms Linear Regression in terms of accuracy, as indicated by lower error values and higher explained variance. The superior model also demonstrates better generalization, with more robust metrics like MedAE, which reduces sensitivity to outliers. Overall, the study highlights the advantages of combining principal component regression with decision tree regression for enhanced predictive accuracy.***

Keywords: Prediction, Machine Learning, Air Quality Index, Principal Component Regression (PCR), Mean Squared Error (MSE), Mean Percentage Error (MPE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Median Absolute Error (MedAE).

I. INTRODUCTION

Air quality is an essential factor influencing public health, environmental quality, and overall well-being. The Air Quality Index (AQI) is a numerical scale used globally to measure and communicate the concentration of various air pollutants, such as particulate matter (PM2.5 and PM10), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3). The AQI provides insight into the cleanliness or pollution levels of the air and highlights potential health risks for the general population. It is divided into various categories to indicate the severity of air pollution. These bands typically range from 0 to 500, with lower values indicating better air quality, and higher values indicating higher levels of pollution and greater health risks.

For example:

- 0-50: Good (air quality is satisfactory)
- 51-100: Moderate (Air quality is acceptable, but could potentially affect the health of some people.)
- 101-150: Unhealthy for Sensitive individuals (May cause health effects for people with respiratory or heart conditions.)
- 151-200:Unhealthy (may affect Individuals with Respiratory or Heart Conditions.)
- 201-300: Severely unhealthy (Health alert: Severe health effects may affect all individuals.)
- 301-500: Dangerous (Health alert due to emergency circumstances)

Given the profound effect of air quality on public health, precise forecasting of AQI is essential for informed decision-making, urban development, and ensuring public safety. Machine learning (ML) has emerged as a powerful tool to predict AQI by analyzing historical data, concentrations. weather conditions, and pollutant By leveraging machine learning algorithms, it is possible to forecast AQI values for future periods, providing valuable information for individuals, governments, and industries to take timely actions to protect health. Machine learning models, including regression analysis, decision trees, random forests, and neural networks, can be trained on large datasets to identify complex patterns between environmental factors and AQI levels. With the increasing availability of environmental data and the rise of advanced computational techniques, predicting AQI with high accuracy has become more feasible.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

II. ENSEMBLE LEARNING

A. *Principal Component Regression with Decision Tree Regression*

Principal Component Regression (PCR) with Decision Tree Regression combines two techniques to improve model performance.

1) Principal Component Analysis (PCA): First, PCA is used to reduce the dimensionality of the data by transforming the features into a smaller set of uncorrelated components. These components account for the majority of the variance in the data.
2) Regression Step: After applying PCA, a regression model is trained on the reduced features (principal components) instead of the original ones. In PCR, this regression can be linear or any other model.
3) Decision Tree Regression: Decision trees are used for regression tasks where the model splits the data into branches based on feature values. A decision tree regression creates a model that predicts a continuous output by making decisions at each split, without requiring a linear relationship
4) Combining PCR with Decision Tree: In this combined approach, PCA is used first to reduce dimensionality and eliminate correlated features. Then, a decision tree regression model is applied on the reduced dataset to predict the target variable.
5) Benefits: PCR helps in handling high-dimensional data efficiently, and the decision tree regression allows for capturing complex, non-linear relationships between the components and the target variable.

In short, PCR simplifies the data, and Decision Tree Regression captures complex patterns in the reduced data, providing a powerful predictive model.

III. EVALUATION METRICS

1) MSE (Mean Squared Error): Measures the average of the squared differences between actual and predicted values. Lower values indicate better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

2) MPE (Mean Percentage Error): Measures the average percentage difference between predicted and actual values. It indicates how much the model's predictions deviate in percentage terms.

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \times 100$$

3) MAPE (Mean Absolute Percentage Error): The average of the absolute percentage differences between actual and predicted values. Lower values indicate better accuracy.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

4) MedAE (Median Absolute Error): Represents the median of the absolute differences between predicted and actual values, offering a robust measure of prediction error.

$$MedAE = \text{median}\left(|y_i - \hat{y}_i|\right)$$

5) Explained Variance: Measures the proportion of variance in the target variable that the model can explain. Higher values suggest the model fits the data well.

$$\text{Explained Variance} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}$$

6) Adjusted R^2: A modified version of R^2 that adjusts for the number of predictors in the model, helping to avoid overfitting by penalizing models with too many variables.

$$AdjustedR^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1}\right)$$

IV. RESULT

Algorithm	MSE	MPE	MAPE	MedAE	Explained Variance	Adjusted R^2
Linear Regression	28.2020	-0.4104	3.8531	2.0088	0.8861	0.8823
Principal Component Regression with Decision Tree Regression	0.8213	0.0129	0.1749	0.0000	0.9967	0.9966

Table 1: Result

Based on the provided performance metrics, the PCR (Principal Component Regression) with Decision Tree Regression model performs better than Linear Regression. The PCR with Decision Tree Regression shows significantly lower values for MSE (0.8213 compared to 28.2020), indicating better model accuracy. Additionally, it has a much lower MAPE (0.1749 vs. 3.8531), which implies better generalization to unseen data. Most notably, the Median Absolute Error (MedAE) is 0.0000 for PCR with Decision Tree, indicating that at least half of its predictions are perfectly accurate, whereas Linear Regression has a MedAE of 2.0088. The Explained Variance and Adjusted R^2 values for PCR with Decision Tree Regression are both near 1 (0.9967 and 0.9966, respectively), indicating that the model explains a significantly larger proportion of the variance in the data compared to Linear Regression (0.8861 and 0.8823). Therefore, PCR with Decision Tree Regression proves to be the more effective algorithm.

V. CONCLUSION

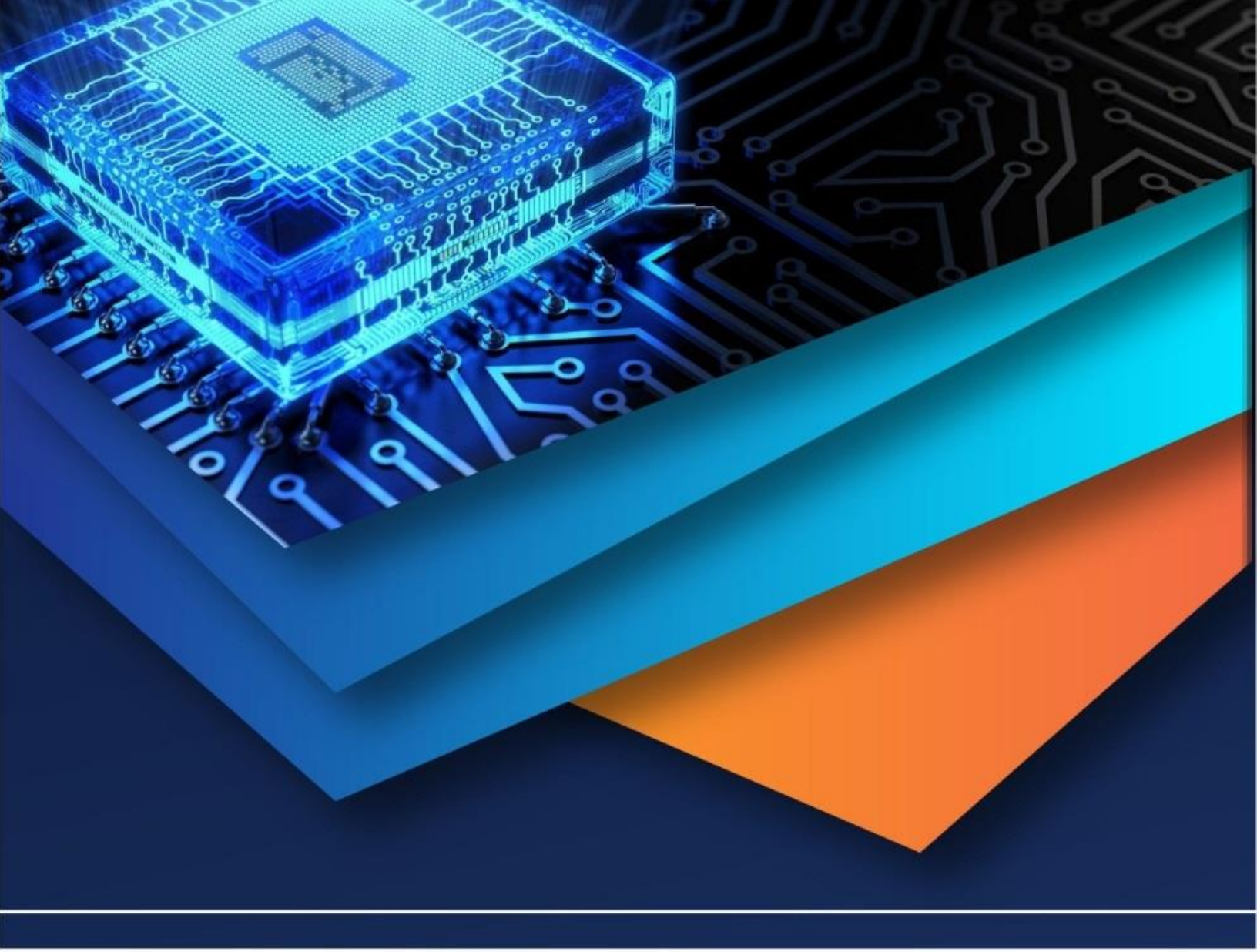
Based on the performance metrics, PCR with Decision Tree Regression outperforms Linear Regression. The model exhibits a significantly lower error across various evaluation measures, including MSE, MPE, and MAPE, which indicates it makes more accurate predictions. The near-perfect MedAE indicates that the model's predictions are highly accurate for at least half of the data points. Additionally, PCR with Decision Tree Regression demonstrates a much higher ability to explain the variance in the data and a stronger overall fit to the data, as indicated by the Explained Variance and Adjusted R^2 values. These factors suggest that PCR with Decision Tree Regression is the more effective algorithm for this particular task.

VI. ACKNOWLEDGEMENT

Thanks to Prof. Pravin Rathod for his help and support throughout this project. His guidance and feedback were very important in making this work better.

REFERENCES

[1] Shorouq Al-Eidi, Fathi Amsaad, Omar Darwish, Yahya Tashtoush, Ali Alqahtani, Niveshitha Niveshitha, “Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques”
[2] R. Sharma, G. Shilimkar, and S. Pisal, ‘‘Air quality prediction by machine learning,’’ Int. J. Sci. Res. Sci. Technol., vol. 8, pp. 486–492, 2021
[3] A. Kumar and P. Goyal, “Forecasting of air quality in Delhi using principal component regression technique,” Atmospheric Pollution Research, vol. 2, no. 4, pp. 436–444, 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY