iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66727

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 I Jan 2025- Available at www.ijraset.com

Air Quality Index Prediction Using Machine Learning Techniques

Chandana H B, Varsha G S, Vaishnavi R, Abhishek K
Department of Electronics and Communication Engineering, JNN College of Engineering, Shimoga

Abstract: ***Air pollution poses a serious threat to public health and the environment, making accurate monitoring and prediction crucial. This research focuses on Air Quality Index (AQI) prediction using machine learning (ML) techniques. We analyze pollutant levels, including PM2.5, NO_2, CO, and O_3, along with meteorological factors such as temperature, humidity, and wind speed. The dataset is collected from cloud-based sources, preprocessed, and split into 80% training and 20% testing sets [1]. Various ML models—Linear Regression, K-Nearest Neighbors (KNN), and Lasso Regression—are applied and evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE). The results demonstrate that KNN outperforms other models in AQI prediction. This study highlights the potential of ML in environmental management, enabling timely interventions for air pollution contro[2].***

Keywords: Air Quality Index, Machine Learning, Linear Regression, KNN, Lasso Regression

I. INTRODUCTION

Air pollution has become a significant global issue, impacting human health and ecosystems. The Air Quality Index (AQI) serves as a standardized metric to quantify pollution levels, helping authorities and individuals make informed decisions. Traditional monitoring techniques, such as manual sampling and sensor-based systems, face limitations in spatial coverage and real-time prediction. Machine learning (ML) techniques offer a promising alternative by analyzing historical and real-time data to predict AQI trends with high accuracy. This study focuses on leveraging ML models for AQI prediction, integrating air pollutant data and meteorological parameters to develop a robust forecasting system. The project aims to enhance real-time air quality monitoring, contributing to environmental management and public health initiatives.

II. MATERIALS AND METHODS

A. Data Collection

The dataset for this study was collected from publicly available cloud-based repositories and governmental agencies such as the Central Pollution Control Board (CPCB). The data consists of Air Pollutants: PM2.5, PM10, NO_2, CO, O_3. Meteorological Factors: Temperature, Humidity, Wind Speed, Atmospheric Pressure. Geographical Scope: Data collected from urban and industrial regions to ensure variability in pollution levels.

B. Data Pre-processing

Is a crucial step to enhance the accuracy and efficiency of machine learning models. The following steps were performed:

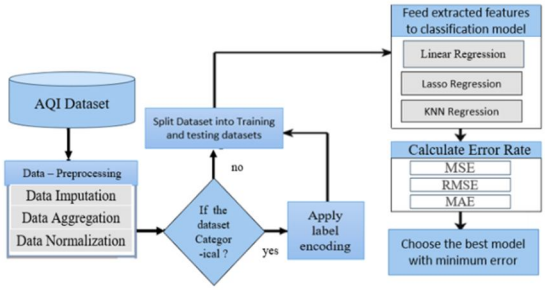


Fig 1: Block diagram of the AQI

C. Data Cleaning

Missing values were handled using imputation techniques (mean/mode imputation). Duplicate and redundant entries were removed.

©IJRASET: All Rights are Reserved

D. Data Normalization

Standardization techniques such as Min-Max Scaling were applied to normalize pollutant concentrations and meteorological parameters Dataset Splitting The dataset was divided into 80% training and 20% testing subsets to train and validate the models.

III. MACHINE LEARNING MODELS

Three different machine learning algorithms were used for AQI prediction:

1) *Linear Regression (LR):* Establishes a direct relationship between pollutants and AQI. Assumes a linear dependence between independent variables and AQI.
2) *Lasso Regression:* Enhances feature selection by eliminating less relevant variables .Reduces model complexity and prevents overfitting.
3) *K-Nearest Neighbors (KNN) Regression:* Predicts AQI by averaging the values of the k-nearest historical data points. Captures complex, non-linear relationships between variables.

A. Model Evaluation

Mean Absolute Error (MAE)

MAE is the arithmetic average of the difference between the ground truth and the predicted values. It can also be de ned as measure of errors between paired observations expressing same phenomenon.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE)

MSE is a common metric used to evaluate the performance of regression models. It quanties the average squared difference between the predicted values and the actual values

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

Root Mean Square Error (RMSE)

RMSE is the square root of the average of the squared difference between the target value and the value predicted by the model. It is square root of mean square error (MSE)

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

B. Implementation Tools

Python Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn. Development Environments: Jupyter Notebook and Spyder.

IV. RESULTS AND DISCUSSION

A. Data Visualization

1) *Pair Grid Analysis:* Showed relationships between pollutants and AQI. .It can show scatter plots between two variables of PM 2.5 and other independent variables respectively It can show histograms or other charts on the diagonal

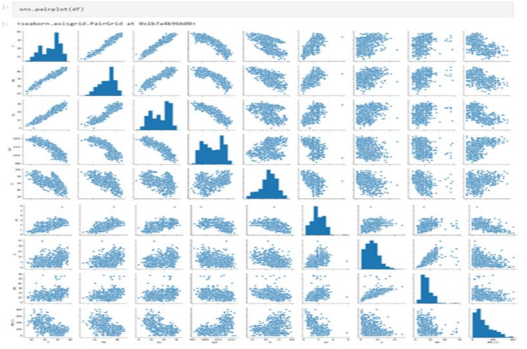


Fig 2: Pairgrid

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

2) *Correlation Heatmap:* Identified significant correlations between meteorological parameters and air quality. In this case, the heatmap visualizes the correlation between different features in a dataset. The code begins by importing the Seaborn library, which is widely used for statistical data visualization. It then computes the correlation matrix of a dataset using the `.corr()` function, which measures the relationship between different numerical features. The correlation values range from-1 to 1, where a value close to 1 indicates a strong positive correlation, a value close to-1 represents a strong negative correlation, and values around 0 suggest little to no correlation between features By analyzing the heatmap, one can identify which features are highly correlated with each other, either positively or negatively. This is useful in various data analysis tasks Similarly, understanding negative correlations can provide insights into inverse relationships between variables

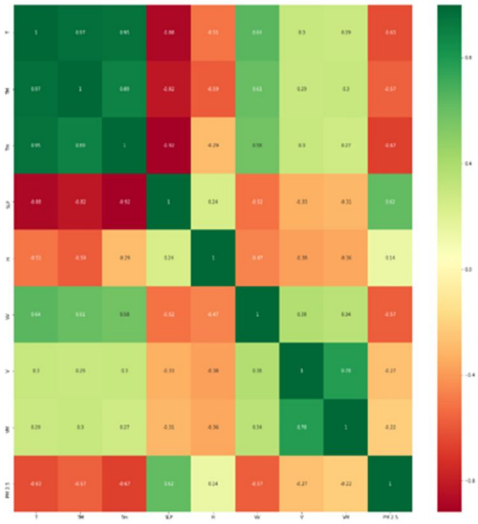


Figure 3: Heat Map

3) *Scatter Plots:* Compared observed vs. predicted AQI values for each model **.**The Output of linear regression : The data is preprocessed and trained with linear regression algorithm to predict the AQI. The figure shows how the linear regression model is configured. scatter plot graph and X-axis and Y-axis are observed AQI value and predicted AQI value respectively. These metrics are statistical criteria that can be used to measure and monitor the performance of a model. As our thesis deals with prediction, we've considered MAE and RMSE as the performance metrics

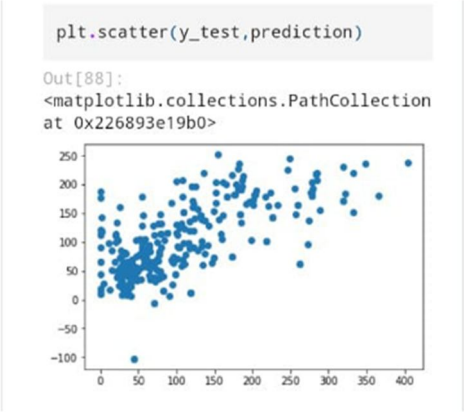


Figure 5: Scatter plot graph for linear regression

Lasso Regression : The data is preprocessed and trained with linear Lasso regression algorithm to predict the AQI. scatter plot graph and X-axis and Y-axis are observed AQI value and predicted AQI value respectively

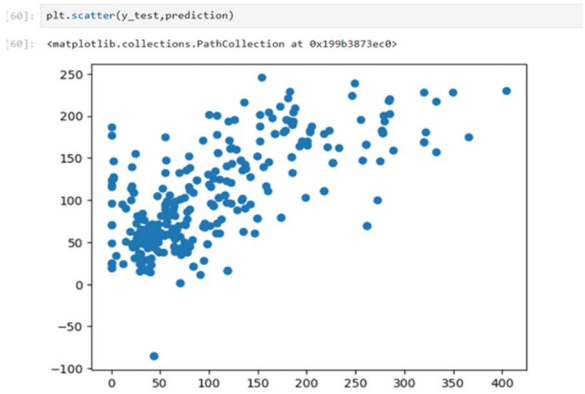


Figure 5: Scatter plot graph for Lasso regression

The KNN Regressor : The data is preprocessed and trained with KNN regressor algorithm to predict the AQI scatter plot graph and X-axis and Y-axis are observed AQI value and predicted AQI value respectively

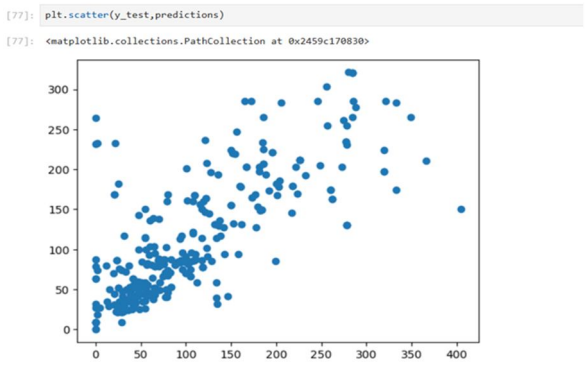


Figure 7: Scatter plot graph for KNN regressor

4) Model Performance Comparison

Algorithm	MAE	MSE	RMSE
Linear Regression	44.8362	3687.5430	60.7251
Lasso Regression	44.508	3627.8109	60.2313
KNN Regression	25.2455	1681.8142	41.0099

Table : Comparison of performance metrics for all models

B. Interpretation of Findings

Seasonal trends influence AQI variations significantly represents a time series visualization of various pollutant concentrations over a specific period, which is likely used in the context of Air Quality Index (AQI) prediction. Each line in the graph corresponds to a di erent parameter or pollutant, such as PM2.5, PM10, CO, NO2, or O3, with their respective concentrations plotted against time

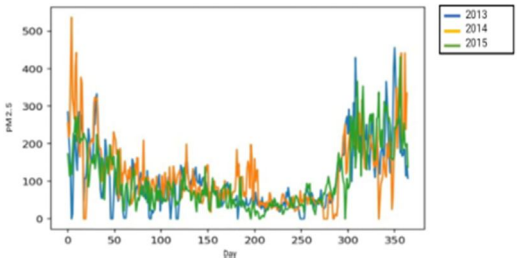


Figure 6: The output of AQI Dataset

KNN provides superior accuracy by capturing non-linear relationships in air pollution data. The visualization is crucial for understanding the variability and trends of pollutants over time, highlighting peaks that can correlate with poor air quality episodes. Such plots are used during data analysis to identify relationships between pollutants and AQI, facilitating model training and feature selection for prediction.

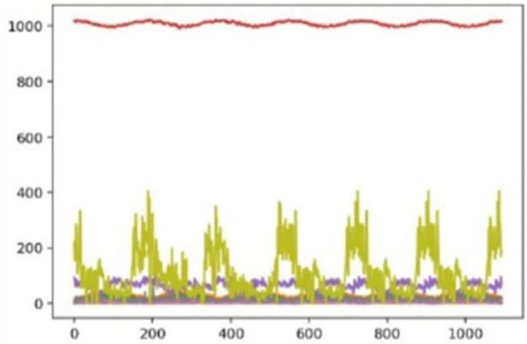


Figure 7: The output of Extract Combine Data

V. CONCLUSION

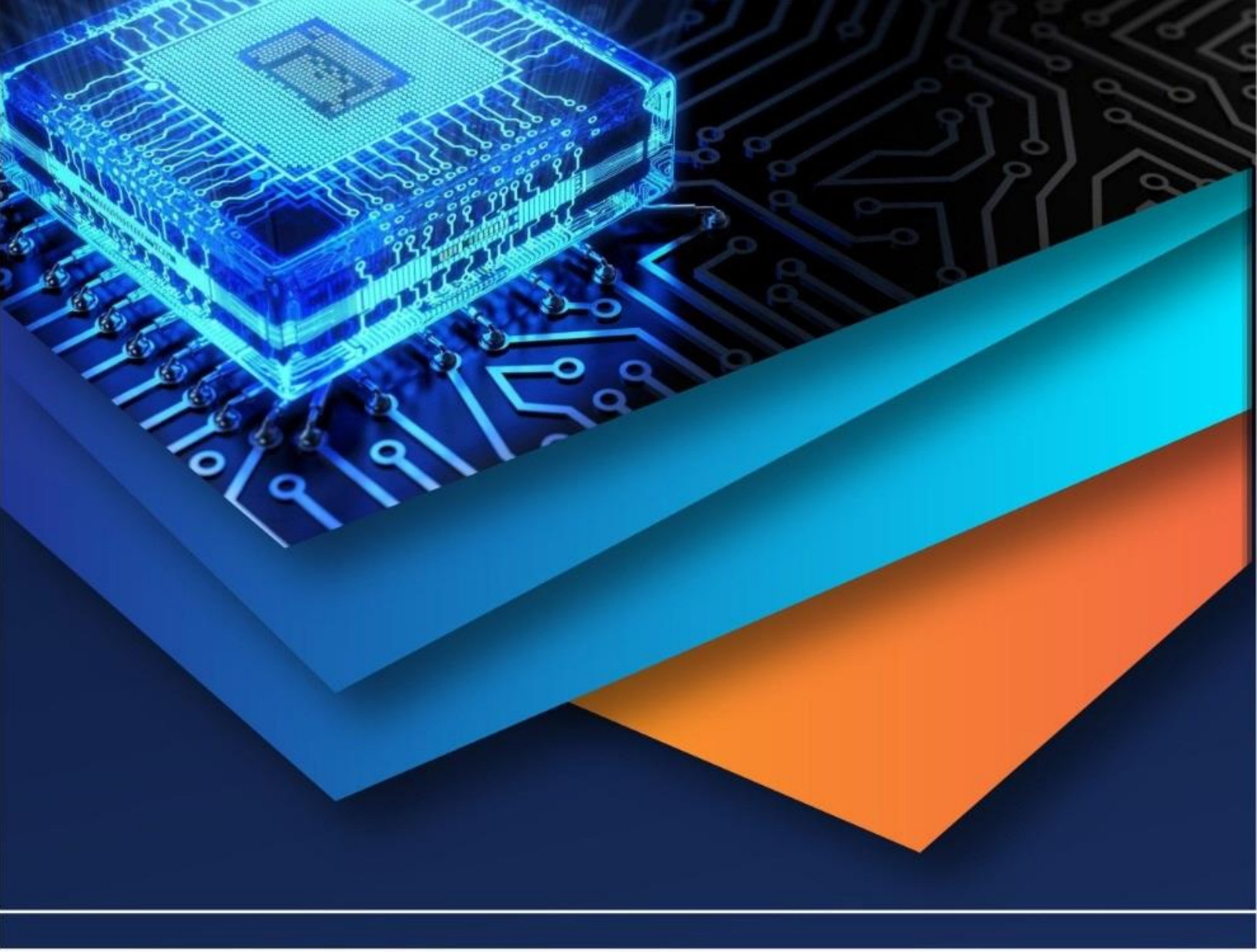
This research successfully implemented machine learning techniques to predict AQI with high accuracy. The study demonstrated that KNN outperforms Linear and Lasso Regression models in predicting air quality trends. The findings emphasize the importance of ML in environmental monitoring and highlight opportunities for real-time AQI forecasting.

Deep Learning Models: Exploring RNNs and Transformers for better time-series forecasting. Real-time Integration: Deploying IoT sensors for live AQI updates Policy Support: Assisting governments in urban planning and pollution control. Global Adaptability: Extending the model for different geographical regions.

REFERENCES

[1] E. Yaacoub, A. Kadri, M. Mushtaha and A. Abudayya, “Air quality monitoring and analysis in Qatar using a wireless sensor network deployment” published in Sensors in 2020, IEEE.
[2] S. Pandya, H. Ghayvat, A. Sur, M. Awais, K. Kotecha et al., “Pollution weather prediction system: Smart outdoor pollution monitoring and prediction for healthy breathing and living” published in 2020
[3] N. Salman, A. H. Kemp, A. Khan and C. Noakes, “Real time wireless sensor network (WSN) based indoor air quality monitoring system” published in IFAC-Papers in 2019
[4] T. Madan, S. Sagar, D. Virmani, “Air Quality Prediction using Machine Learning Algorithms– A Review” Published in
[5] B D Parameshachari, G.M.Siddesh, V.Sridhar, M.Latha, K.N.A.Sattar, and G. Manjula “Prediction and Analysis of Air Quality Index using Machine Learning Algorithms”. Published in 2022.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)