Air Quality Index Prediction Using Ensemble Model

Shraddha Ishwarchandra Ghonsikar[1], Pravin R. Rathod[2]
[1,2]*Deogiri Institute of Engineering and Management Studies*

***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 presents a comparative analysis of machine learning algorithms for predictive modeling, focusing on an ensemble model combining Deep Learning + XGBoost + SHAP Feature Importance and the algorithms from the referred base paper such as Decision Tree Regression and Random Forest Regression. The performance of each algorithm is evaluated using three key metrics: the coefficient of determination (R^2 Score), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics provide insights into the accuracy, consistency, and reliability of the models' predictions. Among the evaluated approaches, the Deep Learning + XGBoost + SHAP ensemble model demonstrates superior overall performance, offering the most accurate and robust predictions across all evaluation criteria.**

Keywords - Prediction, Machine Learning, Air Quality Index, Coefficient of Determination (R^2 Score), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), XGBoost (eXtreme Gradient Boosting), SHAP (SHapley Additive exPlanations).

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.

II. SYSTEM DEVELOPMENT

A. DATASET USED

The dataset utilized in this study consists of a thorough collection of 19231 records, gathered from monitoring stations located across ten different areas within Pune City.The areas covered in the study include BopadiSquare_65,Karve Statue Square_5, Lullanagar_Square_14,Hadapsar_Gadital_01, PMPML_Bus_Depot_Deccan_15,Goodluck Square_Cafe_23,Chitale Bandhu Corner _41, Pune Railway Station_28,Rajashri_ Shahu_Bus_Stand_19.

Dataset Link:

https://www.kaggle.com/datasets/akshman/pune-smartcity-test-dataset

B. SYSTEM ARCHITECTURE

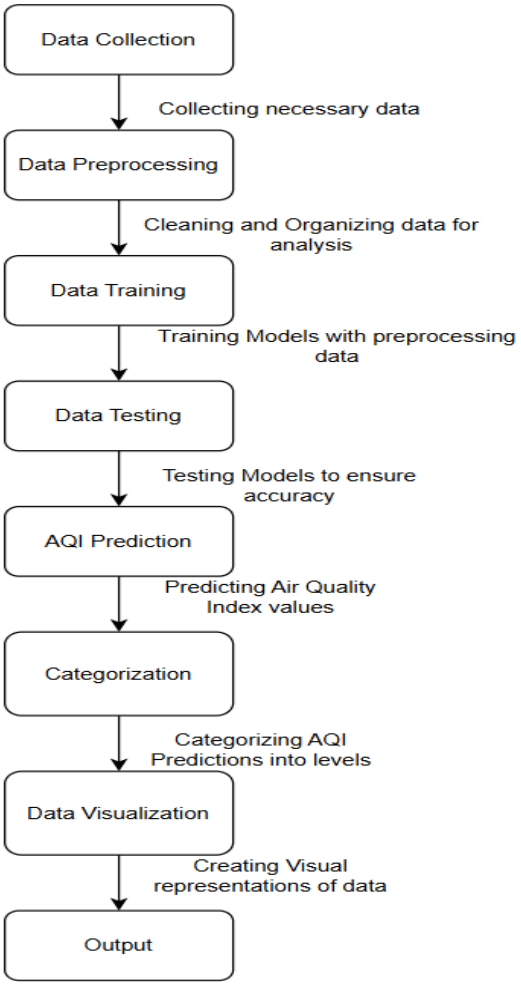

Fig.1. Block Diagram for Air Quality Index Prediction

Explanation:

1. Data Collection: A pre-existing dataset containing air quality and environmental features is used as the input for model development.

2. Data Preprocessing: Raw data is cleaned to handle missing values, outliers, and noise. Features are scaled or transformed to prepare the dataset for training.

3. Data Training: In the data training step, multiple machine learning models are trained on the processed data to learn patterns and accurately predict the Air Quality Index.

4. Data Testing: The models are tested on unseen data, and its performance is evaluated using metrics such as R^2 Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) to ensure prediction reliability.

5. AQI Prediction: The trained model predicts the numerical AQI value based on input pollutant levels and environmental features.

6. Categorization: The predicted AQI value is categorized into standard air quality levels like Good, Moderate, Unhealthy for sensitive groups based on defined AQI ranges.

7. Data Visualization: Graphs, charts visually represent AQI predictions and trends. This helps in analyzing results and communicating insights effectively.

8. Output:

The final output includes AQI prediction results, Algorithms performance, statistics.

II. ALGORITHM USED

Algorithm: Ensemble Model: Deep Learning + XGBoost + SHAP Feature Importance

- Data Collection and Preprocessing: Start by gathering and cleaning dataset, handling missing values and encoding categorical variables. Normalize numerical features for better model performance. Finally, split the data into training, validation, and test sets, ensuring balanced classes if needed.
- Deep Learning Model Training: Train a deep learning model, such as an MLP for tabular data, tuning hyperparameters on the validation set. The model learns complex patterns, and you extract

either its predictions or intermediate layer features for further use.

- Feature Extraction from Deep Learning: Extract learned features (embeddings) from the trained network or use its outputs, then combine these with the original features to create an enriched dataset that captures both raw and abstract information.

- XGBoost Model Training: Train an XGBoost model on this combined feature set. XGBoost efficiently handles tabular data and benefits from the deep features, improving predictive accuracy through its robust tree boosting approach.

- SHAP Feature Importance Analysis: Apply SHAP to interpret the XGBoost model's predictions, identifying which features most influence the output. SHAP visualizations support explain model behavior and emphasize important features for further refinement.

- Feature Selection (Optional): Based on SHAP values, select the top features and retrain the models using this reduced set. This can enhance model speed, simplicity, and generalization by removing irrelevant or noisy features.

- Ensemble Prediction: Combine predictions from the deep learning and XGBoost models through averaging, voting, or stacking to leverage the strengths of both, resulting in better overall performance.

III. EVALUATION METRICS

- R^2 Score (Coefficient of Determination): Measures how well the regression model explains the variance in the target variable. Values range from 0 to 1 (or negative if the model performs less effective than a horizontal line), with higher values indicating better fit.

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

- Mean Absolute Error (MAE): It denotes the average of the absolute errors between the predicted and actual values. It gives an idea of how far predictions are from true values, without considering direction.

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

- Root Mean Squared Error (RMSE): The square root of the average of squared differences between predicted and actual values. It prioritizes larger errors more than MAE and is sensitive to outliers.

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

IV. RESULT

Algorithm	R^2 Score	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
Ensemble Model: Deep Learning + XGBoost + SHAP Feature Importance	0.9929	0.997	7.908
Decision Tree Regression	0.9846	0.7387	8.1667
Random Forest Regression	0.9834	1.0387	8.7042

Table 1: Result

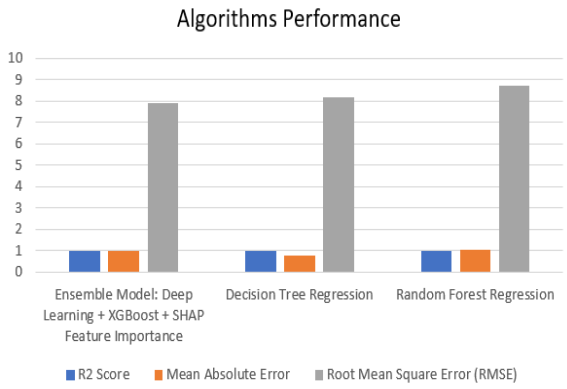

Fig.1: Algorithms Performance

CONCLUSION

The Ensemble model, which combines Deep Learning, XGBoost, and SHAP feature importance,

demonstrates the best overall performance among the evaluated algorithms. It achieves the highest R^2 score of 0.9929, indicating that it explains over 99% of the variance in the target variable. It also preserves a low Mean Absolute Error (MAE) of 0.997 and the lowest Root Mean Squared Error (RMSE) of 7.908, denoting both high accuracy and minimal large prediction errors. This superior performance can be attributed to the ensemble's ability to leverage the complementary strengths of its components: deep learning captures complex, non-linear relationships; XGBoost efficiently handles structured data and stops overfitting through regularization; and SHAP make sure that only the most impactful features are used, improving interpretability and relevance. In comparison, Decision Tree Regression, while having the lowest MAE (0.7387), has a slightly lower R^2 (0.9846) and higher RMSE (8.1667), suggesting less stability in predictions. Random Forest Regression shows the weakest performance overall, with an R^2 Score of 0.9834, MAE of 1.0387, and RMSE of 8.7042. Overall, the Ensemble model offers the most accurate, robust, and generalizable results, making it the most suitable choice for this prediction task.

ACKNOWLEDGMENT

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 Atmospheric Pollution Research, vol. 2, no. 4, pp. 436–444.