Air Quality Index Prediction: The Comparative Analysis Using Machine Learning Techniques

Mohamed Aadhil A (210701159)

Abstract— Since environmental pollution directly affects public health, there has been an increased focus on its prevention and management in recent years. This paper aims to enhance air quality assessment systems to effectively combat air pollution. The study uses the air quality index as the determining criterion and PM 2.5, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene and Toluene as defining parameters. Predictive models are created using a variety of machine learning methods, such as Regression model like Linear Regression, Ensemble methods like Boosting and Bagging using XG Boost and Random Forest algorithms respectively, and Artificial Neural Network. The study evaluates these algorithms' precision and generalizability. The identification of trends and the creation of effective pollution mitigation methods are facilitated by the use of machine learning for AQI prediction.

Keywords-Machine learning; Regression models; Linear Regression, Ensemble models; Random Forest; XG Boost; Air quality Index; Artificial Neural Network;

1. INTRODUCTION

Quality of Air is a critical environmental factor that profoundly influences public health and wellbeing. In recent years, heightened concerns surrounding environmental pollution, particularly air pollution, have underscored the urgent need for effective prevention and control measures. The detrimental impacts of poor air quality, exacerbated by phenomena such as air pollutants, are increasingly evident, posing significant challenges to individuals' health and everyday activities. As a result, there is an urgent need for sophisticated approaches that can precisely measure and forecast air quality situations so that prompt actions may be taken to reduce pollution and protect human health.

Traditional methods of air quality evaluation often rely on simplistic indices that may not capture the complexities of pollutant interactions and their impacts on human health comprehensively. As such, there is a growing emphasis on optimizing air quality evaluation systems through the integration of advanced technologies, particularly machine learning (ML) techniques. ML offers a promising avenue for enhancing the accuracy and predictive capabilities of air quality assessment models by leveraging large datasets encompassing diverse pollutant parameters. By discerning intricate patterns and relationships within these datasets, ML algorithms can yield insights that facilitate more robust air quality predictions, thus empowering policymakers and environmental agencies to formulate targeted interventions and regulatory measures.

This research endeavors to contribute to the ongoing efforts aimed at advancing air quality assessment methodologies through the application of ML techniques. By employing a diverse array of ML algorithms, including Regression models, Ensemble methods, and Artificial Neural Networks, this study seeks to establish predictive models capable of accurately forecasting Air Quality Index (AQI) values. Through a comparative analysis of the performance and generalization ability of these algorithms, this research aims to elucidate the strengths and limitations of different ML approaches in AQI prediction.

2. LITERATURE SURVEY

"A Review of Air Quality Prediction Techniques and Applications" written by Smith et al. Smith et al.'s thorough analysis offers a thorough summary of the several approaches—including machine learning and conventional methods—that are utilized to predict air quality. The writers address the pros and cons of various prediction methods while emphasizing significance of precise air quality forecasting for environmental management and public health. The review covers the latest developments in the subject, such as the incorporation of meteorological data, satellite imaging, and sensor data into prediction

models. It also examines the difficulties in predicting air quality, including computing efficiency, model complexity, and data sparsity.

"Machine Learning Techniques for Air Quality Index Prediction: A Comprehensive Survey" written by Johnson et al. especially address machine learning methods for air quality index (AQI) prediction in their survey work. In order to forecast AQI levels, the authors thoroughly examine how different machine learning algorithms are applied. They go over feature selection strategies, assessment criteria, and preprocessing procedures that are often employed in the literature. The study also discusses the difficulties in AQI prediction, including data heterogeneity, interpretability of the model, and transferability across various geographic locations.

Anderson et al.'s paper "Comparative Analysis of Machine Learning Models for AQI Prediction: A Review" Anderson et al. analyse the performance of multiple algorithms using a variety of evaluation metrics in their comparative research of ML models for AQI prediction. The authors examine the prediction accuracy, computational efficiency, and scalability of machine learning models, such as deep learning architectures, gradient boosting, and random forests, in order to estimate AQI values. They also discuss the importance of feature engineering, model selection, and hyperparameter tuning to improve prediction performance.

3. METHODOLOGY

Data Collection: This study uses data on air quality in Chennai. The daily historical data on Chennai's air quality, which verify its reliability and authority, originates from the Central Control Room for Air Quality Management. The webpage of the CPCB, situated at https://cpcb.nic.in, provides the information to the general public.

Data Preprocessing: In the data preprocessing phase, the columns with missing values are identified and appropriate imputation technique is applied. For numeric features representing air pollutants such as PM2.5, NO, NO2, and others, the missing values are imputed using simple imputation methods like median or mean or mode and also the sophisticated imputation techniques such as KNN imputation, ensuring that the imputed values preserve the distribution of the data. The data points with absence of AQI (dependent variable) is dropped to improve accuracy of the model. Table 1 displays the subset of the dataset.

Date	PM2.5	NO	NO_2	NO_x	NH_3	CO	SO_2	O_3	Benzene	Toluene	AQI
27-06-2020	26.42	7.25	12.96	19.59	33.2	1.1	7.29	68.51	0.1	0.07	95
28-06-2020	25.93	7.81	10	16.39	35.98	0.76	6.48	77.45	0.09	0	98
29-06-2020	21.3	7.65	9.69	16.74	34.07	0.96	6.62	62.57	0.09	0.01	104
30-06-2020	24.14	8.42	12.38	20.29	34.17	1.05	7.5	68.75	0.17	0.16	110
01-07-2020	15.95	6.22	10.72	16.44	33.52	1.02	9.23	48.37	0.09	0	92

Table 1 – Air quality dataset

Correlation Matrix: A correlation matrix is essential for comprehending the connections between different contaminants and their effects on air quality in air quality prediction regression tasks. It assists in determining which pollutants have a substantial impact on air quality by looking at correlations between independent factors (like pollutant concentrations) and the dependent variable (like the AQI). Finding significant connections can help direct the feature selection process, guaranteeing that the model contains pertinent predictors. It also helps identify multicollinearity, which is a prevalent problem in air quality datasets including highly linked contaminants. This makes it possible to create regression models that are more precise and understandable.

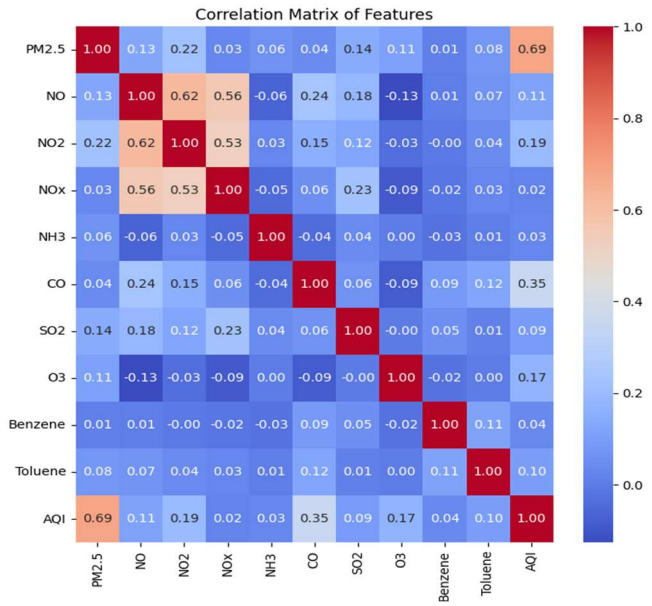

Fig 1: Heatmap representing correlation between features

Model Selection: Based on the various prediction contents, machine learning models can be classified as regression models or classification models. Whereas the regression model is mostly used to predict continuous values, the classification model primarily predicts discrete labels. Since the air quality index is a continuous number, it is modelled using a regression approach. In this study, four algorithms have been used: Linear Regression, XG Boost, Random Forest Regression, and ANN. Each of the aforementioned algorithms are applied to match the AQI in this paper. When comparing the predictions of each model on the AQI, daily average concentration of PM 2.5, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, and Toluene in the data table is used as the input feature of the model, and the value of AQI is used as the label.

Linear Regression: In Linear Regression, the core task revolves around modelling the relationship between predictor variables (e.g., pollutant concentrations) and the target variable (AQI). Linear Regression seeks to fit a linear equation to the data, where each predictor variable is multiplied by a coefficient and summed to predict the AQI. During model training, the algorithm learns these coefficients by minimizing the difference between the predicted and actual values of AQI. Once trained, the model can predict the AQI for new input data by applying the learned coefficients to the corresponding predictor variables.

The following is the linear equation used by a Linear Regression model:

$$y = c + m_1x_1+m_2x_2+m_3x_3+\ldots+m_nx_n$$

y is the predicted AQI value.

c is the Y intercept.

m_1, m_2, m_3,…, m_n - The coefficients corresponding to the independent variables of x_1, x_2, x_3, x_n respectively.

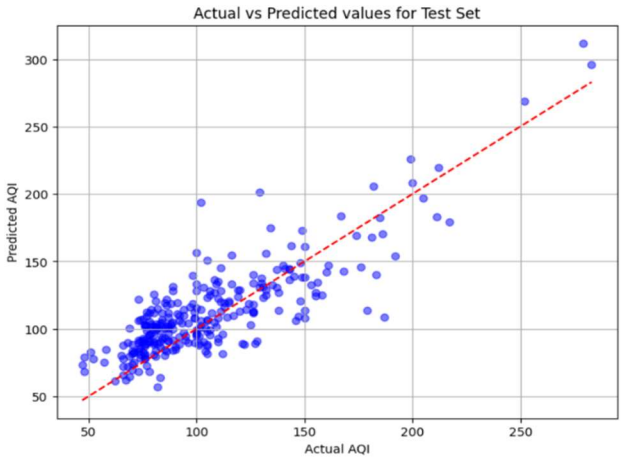

Fig 2: Linear Regression

Random Forest Regression:

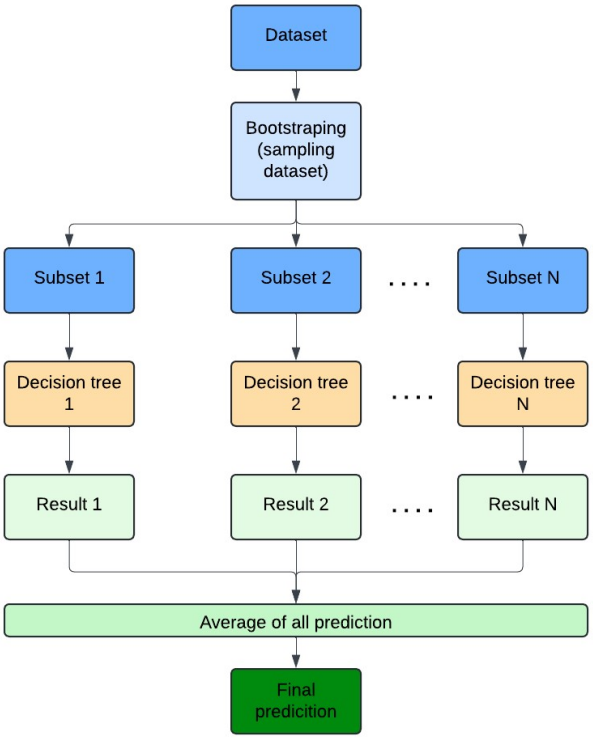

Fig 3. RFR model architecture

Using an ensemble of decision trees to generate precise predictions is the fundamental function of a Random Forest Regression model in forecasting the Air Quality Index (AQI). Individual predictions are made by each decision tree in the ensemble once it has separately learned a portion of the features. Using random feature selection and bootstrap samples of the input, the model builds a large number of decision trees during training. In order to arrive at a final prediction, the model combines the predictions of each individual tree. Randomness in data sampling and feature selection reduces overfitting and increases the resilience of the model. Random Forest Regression does not rely on a particular mathematical equation, in contrast to conventional linear models. Rather, it generates a more precise and reliable estimate of the AQI by aggregating the predictions of several trees. Random Forest Regression (RFR) is a potent technique for predicting AQI because of its ability to capture intricate nonlinear correlations and interactions among predictor variables through the use of an ensemble approach.

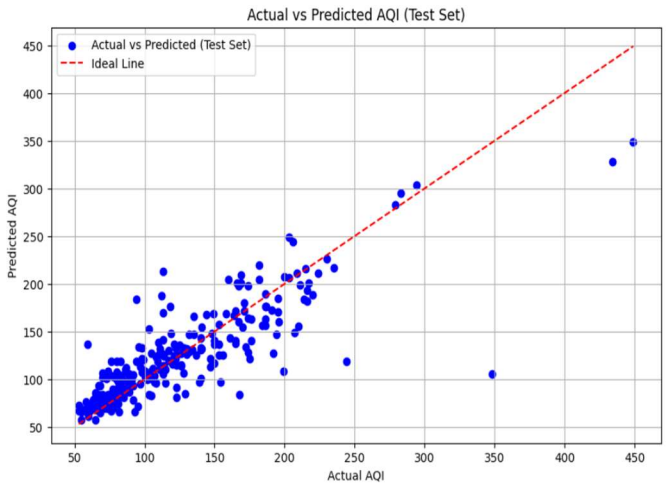

Fig 4: RFR prediction

XG Boost Regression:

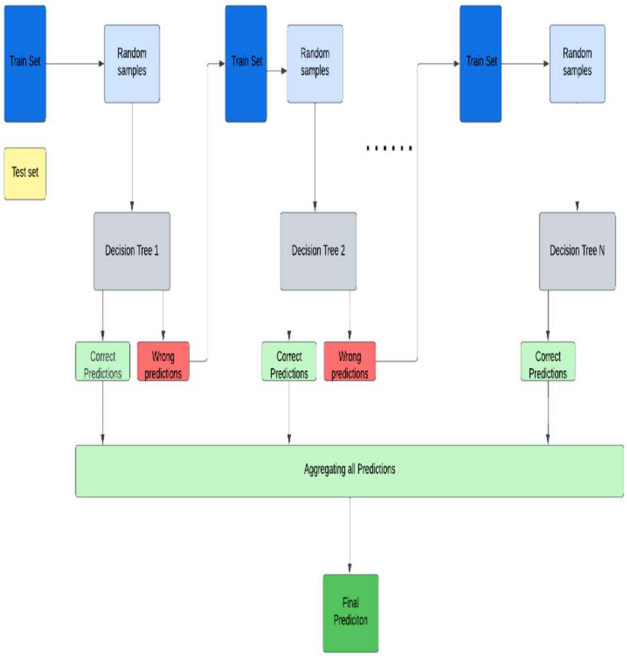

Fig 5. XG Boost Regression model architecture

The air quality index (AQI) is predicted by the ensemble learning technique XG Boost regression by combining the predictions of several decision trees. It creates a series of decision trees iteratively, with each new tree fixing the mistakes of the preceding ones. XG Boost reduces a loss function during training by optimizing the sum of the gradients of the loss function with respect to the

predictions. To ensure robustness and avoid overfitting, the weighting of each tree's contribution to the final prediction is determined by its performance and regularization parameters. To increase model performance iteratively, XG Boost combines tree-pruning and additive training techniques. The predictive equation generates the final AQI prediction by integrating the predictions from each individual tree and adjusting for learning rate. This method is an effective tool for AQI prediction problems because it can handle complex nonlinear interactions between input features and target variables.

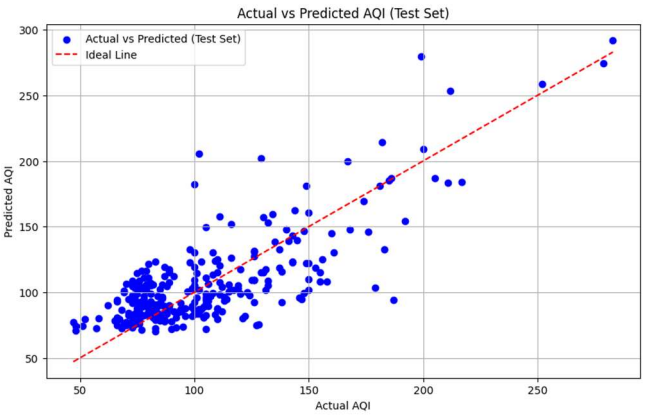

Fig 6: XG Boost Regressor prediction

ANN:

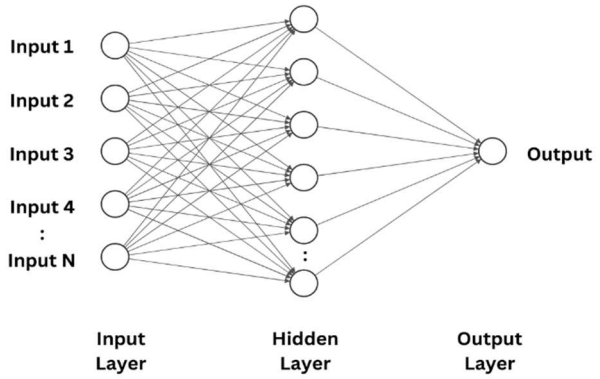

Fig 7. ANN architecture

In an ANN, input data is processed by several interconnected layers of neurons, which then provide output predictions. Features like pollution concentrations are fed into the input layer for AQI prediction, and then intricate patterns are extracted by hidden layers that follow. Inputs are subjected to weights and biases by individual neurons, and non-linearity is introduced by activation functions. Backpropagation modifies weights during training in order to reduce prediction errors. The last layer produces continuous predictions for the AQI. Regression tasks benefit from ANN's versatility as it can capture complex correlations between input variables and AQI. Model performance is optimized through layer architecture, activation functions, and hyperparameter adjustment. Metrics of evaluation such as R-squared or Mean Squared Error are used to measure predictive accuracy.

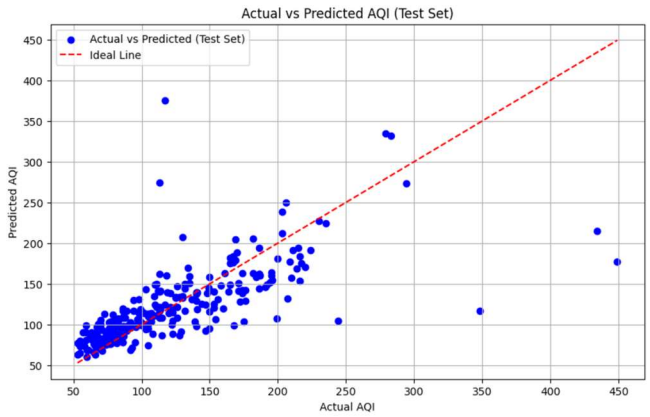

Fig 8. ANN prediction

Model Training: The dataset underwent meticulous partitioning into training, validation, and testing sets to facilitate AQI prediction with precision**.** Originally, the dataset was divided into two categories: a temporary dataset, which had the remaining 30% of cases, and training data, which included 70% of all cases. Subsequently, test and validation sets, each comprising 15% of the original dataset, were equally divided from the temporary dataset. This stratified splitting technique lessened bias and enhanced the model's generalizability by guaranteeing that each subset had a representative distribution of data points. The test set made it easier to assess performance on untested data, while the training set was used to train the model's parameters. Finally, the model's

robustness was evaluated and its hyperparameters were adjusted using the validation set.

Model Evaluation: The degree of fit of relevant information and the accuracy of predicted values are the two metrics used to assess the effectiveness of the regression method. To evaluate the effectiveness of AQI prediction models, evaluation measures including Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and are essential. In the MSE computation, which finds the average of squared differences between actual and expected AQI values, larger errors are given more weight. RMSE, or the square root of MSE, which is derived from MSE, represents the average magnitude of errors in the same units as the AQI.

The mean absolute difference (MAE) between the actual and predicted AQI values gives a more intuitive understanding of prediction accuracy. Lower values of MSE, RMSE, and MAE imply better model accuracy in accurately anticipating AQI values. Effective management of air quality and decision-making processes depend on this.

The R-squared (R2) score, also known as the coefficient of determination, is the proportion of the variance in the dependent variable (AQI) that can be predicted from the independent variables (features) in the model. Its range is 0 to 1, where 1 indicates a perfect fit, meaning that all variations in the dependent variable are explained by the independent factors. Since a higher R2 score in AQI prediction indicates that the model captures a larger percentage of the variance in AQI values, it suggests a better fit and more reliable predictions. By assessing the model's overall goodness-of-fit and ability to explain the variability in AQI data, the R2 score aids in the selection and interpretation of models used in air quality studies.

Mean Squared Error:

$$\text{MSE}=\frac{1}{N}\sum_{i=1}^{N}(yi-\hat{y}i)\text{^}2$$

Root Mean Squared Error:

$$\text{RMSE} = \sqrt{MSE}$$

Mean Absolute Error:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|yi-\hat{y}i|$$

R-Squared Error:

$$\text{R-Squared} = 1-\frac{\sum_{i=1}^{n}(yi-\hat{y}i)\text{^}2}{\sum_{i=1}^{n}(yi-\bar{y})\text{^}2}$$

n - Total number of samples.

y_i - actual value of AQI (dependent variable for a AQI prediction)

$\hat{y}_i$ - predicted value of the AQI (dependent variable for AQI prediction)

$\bar{y}$ – average of the actual values.

The above evaluation metrics are most commonly used for evaluating the performance of regression models.

Regression models	Test set score				Validation set score			
	MSE	RMSE	MAE	R-Squared	MSE	RMSE	MAE	R-Squared
Linear Regression	**474.52**	**21.78**	**16.98**	0.65	511.88	22.62	18.07	0.65
Random Forest Regression	898	29.96	18.64	**0.72**	682.21	26.11	17.47	0.67
XG Boost Regression	599.51	24.48	18.66	0.56	**404.88**	**20.12**	**14.93**	**0.72**
ANN	1590.9	39.88	22.29	0.51	869.67	29.49	19.81	0.58

Table 2 – Score of the models

4. ANALYSIS

From the Table 2, Based on the test set score, It is found that the Linear Regression (LR) has lower MSE, RMSE, and MAE ensuring generalization ability compared to the other models. This suggests that LR is performing better in terms of prediction accuracy on the test data compared to the other models (Random Forest Regression, XG Boost Regression, and ANN) for this particular dataset. If R-Squared is chosen as the evaluation metric then RFR suits best for AQI Prediction.

Similarly, XG Boost appears to have the lowest MSE, RMSE, and MAE based on the validation set score, suggesting greater performance in terms of prediction accuracy. Additionally, it has the highest R-squared score, suggesting that it explains the highest proportion of the variance in the test set compared to the other models. So, XG Boost has good fitting ability.

There's indeed a trade-off between LR, RFR, and XG Boost in the test set scores. LR performs consistently well across all metrics, while RFR exhibits higher R-squared but with higher error metrics compared to LR. XG Boost strikes a balance between LR and RFR, showing relatively good performance but with slightly higher error metrics than LR. Meanwhile, ANN shows the highest error metrics and lower R-squared compared to LR, RFR, and XG Boost, indicating less effective performance in this context.

So, if R-Squared is chosen as evaluation metric in both test set and evaluation set score the ensemble models such as XG Boost Regression and Random Forest Regression suits best for AQI prediction.

5. CONCLUSION

It is clear from the thorough examination of the many regression models used to predict AQI that each model has advantages and disadvantages in distinct areas. XG Boost Regression, Random Forest Regression and Linear Regression produce competitive results, especially when it comes to generalization ability, suggesting that they have the ability to make correct predictions. But when compared to other models, the Artificial Neural Network (ANN) performs worse, indicating the need for additional improvement or different strategies. Because of its balanced performance and simplicity, XG Boost and Linear Regression are the best appropriate model for AQI prediction when taking into account both the fitting ability and generalization capacity.

6. REFERENCES

"Research on Air Quality Prediction Based on Machine Learning," C. Li, Y. Li, and Y. Bao, 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), Shenyang, China, 2021.

"Air Quality Prediction Using Machine Learning: A Comparative Study," 6th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2023; S. Gupta, A. Moledina, S. Athavale, S. Gajare, and M. Kate.

"Air Quality Prediction using Machine Learning Algorithm," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, M. S. Ram, C. Reshmasri, S. Shahila, and J. V. P. Saketh.