WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Forecasting of Air Quality Index Using Machine Learning Techniques for Chennai City

Meenakshi CT[1], Dr Jerline Amutha A[2]

[1]Student, PG Department of Computer Science and Technology, Women's Christian College, Chennai-32, Tamil Nadu, India
[2]Associate Professor, PG Department of Computer Science and Technology, Women's Christian College, Chenna-06, Tamil Nadu, India

A B S T R A C T

Air Quality is an important factor influencing health and environmental standards for people and other organisms. The Air Quality Index (AQI) is a measure used to evaluate air pollution levels and their impact on public health. More reliable AQI forecasts are required due to increased industrialization and a growing number of vehicles that have exacerbated air pollution issues. The objective of this research is to predict the Air Quality Index (AQI) using Machine Learning Techniques for 9 different monitoring stations in Chennai. The specific purpose of this paper is to calculate the average AQI for a certain month in 2023 by analyzing historical data and predict the average AQI for the same month in 2024. The AQI value is then divided into categories such as good to unhealthy, etc., accompanied by health precautions that should be taken in case of exposure. The research paper employs secondary data collected from the Central Pollution Control Board, Government of India for the periods 2020-2023. This paper seeks to compare and evaluate SVR, Random Forest Regression, and XG Boost models for forecasting and predicting Air Quality Index. To improve the accuracy of the predictions, stacking techniques within an ensemble method are employed, which combine the strengths of individual models. Collectively, the findings of this study assist in enhancing air quality predictions for combating pollution and reducing its impact on human health.

Keywords: Air Quality Index (AQI), Air pollution, Machine learning, Support Vector Regression (SVR), Random Forest Regression, XG Boost.

1. INTRODUCTION

Air quality has emerged as a pressing concern in urban settings, where rising pollution levels pose serious threats to public

health and environmental sustainability. Air pollution from industry and vehicles causes global warming, acid rain, species extinction, and health issues. Tiny particles, as small as 0.01mm, can lead to severe illnesses like cardiovascular disease [2].

The Air Quality Index (AQI) measures pollutants like PM2.5, PM10, CO, NO, NO2, SO2, and O3, providing crucial data for public and policymaker action. In densely populated cities like Chennai, rapid urbanization has worsened air quality. A recent study in "The Lancet Planetary Health" found that poor air quality led to about 2,900 annual deaths in Chennai from 2008 to 2018, highlighting the need for accurate AQI predictions to protect public health. The Air Quality Index (AQI) ranges from

0 to 500, with six levels from "Good" (0-50) to "Hazardous" (301-500). It can be calculated using specific pollutant concentrations or predicted using machine learning models.

This research aims to develop a predictive model for forecasting the AQI in Chennai using machine learning techniques. By analyzing historical air quality data and predicting levels of the Air Quality Index and categorizing them, the research covers the period from 2020 to 2023 and includes data from nine air quality monitoring stations in Chennai: Alandur Bus Depot, Manali Village, Velachery, Royapuram, Gandhi Nagar, Arumbakkam, Kodungaiyur, Manali Chennai, and Perungudi. The findings of this research will enhance AQI monitoring and management by demonstrating the importance of predictive analytics in addressing urban air quality challenges. This research aims to inform evidence-based policies and support effective pollution control strategies tailored to Chennai, helping to mitigate air pollution and protect public health.

2. LITERATURE REVIEW

In [2],[3] Machine learning (ML) algorithms play a key role in predicting the Air Quality Index (AQI). They offer useful insights to manage environmental health in cities struggling with pollution. AQI measures air quality based on pollutants such as CO2, NO2, SO2, and particulate matter (PM2.5 and PM10), and plays a critical role in assessing the impact of air pollution on public health. In [4],[6],[8],[10] Support Vector Regression and Random Forest models have been widely used to predict AQI.

In [7], Linear Regression and Ridge Regression are favored for their simplicity and reliability. In [4], Support Vector Regression (SVR), particularly with a Radial Basis Function (RBF) kernel, has proven highly effective in forecasting pollutant levels and AQI, achieving an accuracy of 93.4%. In [8] and [10], Decision Trees and Random Forest algorithms have also been effective in capturing the non-linear interactions between pollutants and their impact on AQI. In [6], [8], and [10], XG Boost frequently emerges as the best performer, offering the highest accuracy and overall predictive capability in AQI prediction.

In [3] and [8], while Neural Networks are powerful, they require extensive tuning to achieve optimal results. In [6], [8], and [10], boosting models, particularly XG Boost, are often preferred for their robustness and ability to handle complex datasets, making them highly suitable for AQI forecasting. In specific case studies, cities like Pune and Mumbai have been identified as needing immediate attention due to rising SO2 levels.

In [3], models like AR and ARIMA have been effective in predicting SO2 levels, with identified safe thresholds at 0.20 ppm (1 hour) and 0.08 ppm (24 hours). In New Delhi, the effectiveness of SVR with an RBF kernel suggests that models incorporating non-linear kernels may offer superior performance in highly polluted environments. In [3], [8], and [9], the application of Random Forest, XG Boost, and SVR models has demonstrated that proper data preprocessing and feature selection are critical for achieving high predictive accuracy. In [3], [8], and [9], the inclusion of meteorological data, such as temperature, wind speed, and humidity, has also been shown to significantly enhance model performance.

In practical applications, the integration of satellite imagery, deep learning, and cloud-based platforms could further refine AQI predictions and enhance their practical utility. In [6], [9], and [11], future research is expected to explore these areas, aiming for real-time, localized predictions that could better inform public health policies and interventions. In conclusion, despite these advancements, challenges remain, particularly concerning data quality and the need for continuous updates. In [11], the studies emphasize the importance of clean, balanced datasets, suggesting that future research should focus on incorporating more complex data sources, such as satellite and sensor data, to improve the precision and reliability of AQI predictions.

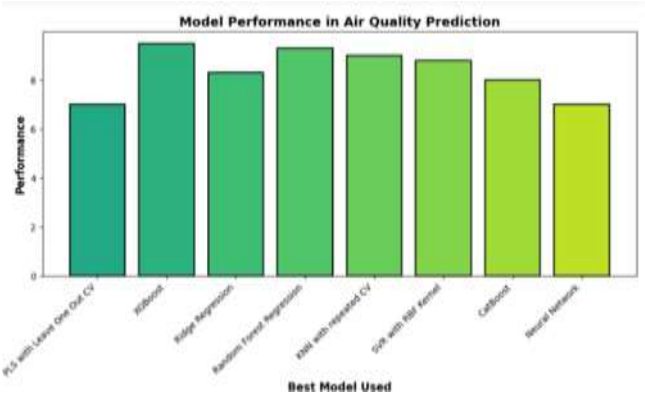

Fig. 1 Graphical Representation of models and performance mentioned in Literature Review

Figure 1 illustrates the performance of various models used in air quality prediction research papers. The x-axis represents the different models, including PLS with Leave-One-Out Cross-Validation, XG Boost, Ridge Regression, Random Forest Regression, KNN with Repeated Cross-Validation, SVM with RBF Kernel, Cat Boost, and Neural Network. The y-axis indicates the performance of these models, likely evaluated using metrics such as accuracy or RMSE, with higher values signifying better performance.

3. METHODOLOGY

3.1 DATA COLLECTION AND ANALYSIS

The dataset for this research was sourced from the Central Pollution Control Board (CPCB) and is publicly available through their official repository [1]. It covers the period from 2020 to 2023 and includes data from nine air quality monitoring stations in Chennai. Initially dataset available separately for each station, then data was consolidated into a single dataset containing 10,810 rows and 25 features. Key features include Timestamp, PM2.5, PM10, NO2, NOx, NH3, SO2, CO, Ozone, Benzene, Toluene, Xylene, Eth- Benzene, MP-Xylene, AT (°C), RH, WS (m/s), WD (deg), RF (mm), TOT-RF, BP (mmHg), SR (W/m²), VWS (m/s), and Station. Among these, the pollutants PM2.5, PM10, NO2, NOx, NH3, SO2, CO, and Ozone were used to compute the Air Quality Index (AQI), which was added as a new column and designated as the target variable. The Timestamp feature was processed to extract the month and year for further analysis, and after merging the data and adding these columns, the dataset was cleaned to address noisy data, including null values, missing entries, and outliers, ensuring it was ready for modelling and analysis.

3.2 AQI CALCULATION

Air Quality Index (AQI) calculation methods vary significantly across different countries, each designed to adhere to local environmental standards and health guidelines. For example, China's AQI, based on the National Ambient Air Quality Standards of China (NAAQS-1996), differs from the AQI calculation methods used by the U.S. Environmental Protection Agency (EPA, 1994) and India's NAAQS-Dependent Air Quality Index [10]. These differences reflect the unique environmental conditions and public health priorities of each region.

In this study, we adopted the AQI calculation method outlined by the National Ambient Air Quality Standards (NAAQS) of India. According to NAAQS-2012, the AQI is calculated based on the concentrations of six key pollutants: PM10, PM2.5, SO_2, O_3, NO_2, and CO. The calculation is based on the maximum 24-hour concentration of each pollutant. The AQI for each individual pollutant is determined using the following equation:

$$AQI = \frac{AQI_h - AQI_l}{BP_h - BP_l} \times (C_q - BP_l) + AQI_l \quad \text{(Eq 1)}$$

Where:

- **AQIH**: AQI value corresponding to the upper concentration breakpoint of the pollutant.
- **AQIL**: AQI value corresponding to the lower concentration breakpoint of the pollutant.
- **BPH**: Upper concentration breakpoint of the pollutant.
- **BPL**: Lower concentration breakpoint of the pollutant.
- **CQ**: Measured concentration of the pollutant.

The Eq(1) computes the AQI by first determining the position of the actual pollutant concentration (CQ) between the lower (BP) and upper (BPH) breakpoints. This is done by subtracting the lower breakpoint (BPL) from the actual concentration (CQ) and dividing the result by the range between the upper and lower concentration breakpoints (BPH – BPL). The resulting value is scaled by the difference between the upper and lower AQI values (AQIH – AQIL). Finally, the scaled value is added to the lower AQI value (AQIL), yielding the final AQI for the measured pollutant concentration.

3.3 DATA PREPROCESSING

The preprocessing phase of the dataset was carried out systematically to maintain data integrity and enhance its suitability for AQI prediction.

1. **Column Selection**: Irrelevant columns such as 'Xylene', 'O Xylene', and other weather-related attributes were dropped to focus on key pollutant concentrations and station information.
2. **Handling Missing Values**: Missing values were replaced with NaN, followed by forward and backward filling to ensure continuity. Any remaining missing values were filled with appropriate means or interpolations. Figure 3.3.1 illustrates the number of missing values per column.

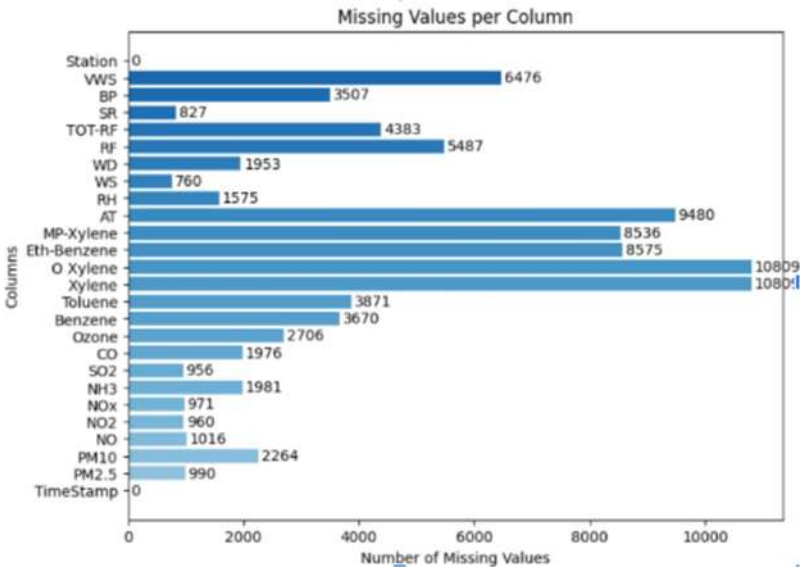

Fig. 2 Graphical Representation of Null Values

3. **Column Standardization**: Column names were standardized by removing special characters and spaces for easier data manipulation.
4. **Outlier Detection**: Outliers were detected using the Interquartile Range (IQR) method for numeric columns. Despite the detection of outliers, they were retained to capture extreme pollution levels.
5. **Label Encoding**: The categorical 'Station' column was encoded into numerical values using label encoding to make it compatible with machine learning models.

6. **Feature Selection**: Recursive Feature Elimination (RFE) with a Random Forest Regressor was used to identify key predictors such as 'PM2.5 (µg/m³)', 'PM10 (µg/m³)', 'NO2 (µg/m³)', 'SO2 (µg/m³)', 'CO (mg/m³)', 'Ozone (µg/m³)', and important temporal variables like 'Station', 'Month', and 'Year'.
7. **Feature Scaling**: Robust scaling was applied to the selected numerical features to ensure consistency and reduce the influence of outliers.

Finally, the pre-processed dataset was saved in Excel format for further analysis. The data was then split into training (80%) and testing (20%) sets using the train_test_split method, with the training set further divided into training (60%) and validation (20%) subsets for comprehensive evaluation of model performance. This thorough preprocessing ensured that the dataset was optimized for accurate and reliable AQI prediction, forming a strong foundation for subsequent model training and evaluation.

Presented below are the first 20 records from the pre-processed dataset, showcasing the refined data utilized for the analysis.

PM2.5(ug	PM10(ug/	NO2(ug/n	SO2(ug/m	CO(mg/m	Ozone(ug	Station	Month	Year	AQI
0.209243	3.192156	0.512302	-0.32535	-0.38065	0.006969	-0.8	-1	-1	87.44635
0.615319	3.192156	0.373462	-0.18662	-0.3502	-0.27381	-0.8	-1	-1	67.78894
0.922978	3.192156	0.407733	-0.0577	0.091356	0.308102	-0.8	-1	-1	103.9196
0.473684	3.192156	-0.00176	-0.00859	-0.31975	0.606104	-0.8	-1	-1	51.15578
0.728284	3.192156	0.118629	0.311848	-0.59381	1.054933	-0.8	-1	-1	81.05528
-0.01926	3.192156	0.27065	0.103131	-0.27407	0.072206	-0.8	-1	-1	75.98712
0.155327	3.192156	0.458699	-0.08103	-0.50246	0.146315	-0.8	-1	-1	84.74249
0.642276	3.192156	1.531634	0.033149	-0.48723	0.059159	-0.8	-1	-1	70.95477
0.629012	3.192156	1.15993	-0.30448	-0.22839	-0.04104	-0.8	-1	-1	69.39698
0.801883	3.192156	1.120387	-0.16329	-0.39588	0.113958	-0.8	-1	-1	89.69849
-0.27	3.192156	1.084359	-0.31308	-0.33497	0.206855	-0.8	-1	-1	63.41202
1.209243	3.192156	1.176626	0.173112	-0.31975	0.316975	-0.8	-1	-1	137.5377
1.620881	3.192156	1.155536	0.597913	-0.22839	0.56644	-0.8	-1	-1	61.1275
3.824561	3.192156	1.458699	0.569675	-0.07613	0.524688	-0.8	-1	-1	142.529
3.824561	3.192156	1.458699	0.569675	-0.07613	0.524688	-0.8	-1	-1	142.529

Fig. 3 First 10 rows from Preprocessed dataset

3.4 MODELLING WITH VARIOUS ALGORITHMS

This research investigates the performance of three machine learning algorithms in predicting Air Quality Index (AQI) values: Support Vector Regression (SVR), Random Forest Regression (RFR), and XG Boost. These algorithms were selected for their proficiency in handling non-linear relationships and delivering high predictive accuracy in complex datasets.

3.4.1 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a robust supervised learning technique utilized for regression tasks, based on the principles of Support Vector Machines (SVM). It seeks to identify a function that predicts the target variable while maintaining a specified margin of error, allowing the model to disregard insignificant deviations. SVR employs the kernel trick, enabling it to capture complex, non-linear relationships by mapping data into higher-dimensional spaces. Its focus on support vectors key data points closest to the regression boundary enhances generalization and reduces the risk of overfitting. This makes SVR particularly effective for various applications, including environmental monitoring and financial forecasting.

In this paper Support Vector Regression (SVR) with the RBF kernel was employed to model the non-linear relationships between air pollution features and the AQI. The RBF kernel helps capture intricate patterns that linear models may miss, making it suitable for AQI prediction where pollutant concentrations and AQI are often non-linear. However, despite its potential, the model achieved an R^2 score of 45% of the variance in AQI. This suggests that while SVR captures some non-linear patterns, it may struggle with AQI's complexity and high variability.

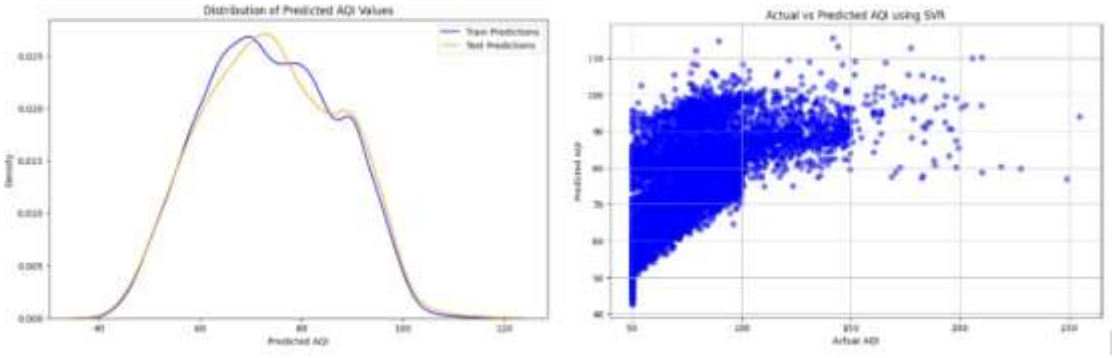

Fig. 4 (a): Density plot for SVR (b): Scatter plot for SVR

Density plot shows that the predicted AQI distribution closely follows the actual AQI, though deviations in the tails indicate the SVR model's challenges in predicting extreme values. In Scatter plot, the scatter plot illustrates that while SVR performs well at lower AQI levels, the error margin increases as the actual AQI values rise. This widening spread suggests the model struggles with higher pollution levels, possibly due to the complexity of these extreme cases. The overall R^2 score of 0.45 reflects the model's partial ability to capture AQI variability.

3.4.2 Random Forest Regression (RFR)

Random Forest Regression (RF) is a machine learning technique that builds a collection of decision trees to predict continuous outcomes. By aggregating the predictions of these multiple trees, RF reduces correlation among individual predictions, thereby minimizing overall error. This method is especially effective when working with large datasets, as it can handle high-dimensional feature spaces efficiently.

Two crucial factors that influence the performance of a Random Forest model are the number of trees in the collection and the number of random variables considered at each split. By averaging the predictions from various trees, RF effectively reduces the risk of overfitting, resulting in a more generalized model. This approach capitalizes on the strengths of diverse decision trees, ensuring that the final prediction reflects a robust consensus. As a result, Random Forest Regression is a powerful and adaptable tool for various regression tasks in data science.

In this research Random Forest Regressor (RFR) was selected for its robust performance in handling large datasets with numerous features, making it highly suitable for AQI prediction. The model was trained and evaluated using the same procedure as SVR, but RFR achieved a significantly better performance, with an **R^2 score of 0.95** of the variance in AQI values. This improvement is attributed to the ensemble learning approach of RFR, which combines multiple decision trees to reduce overfitting and enhance the model's ability to generalize across different data points.

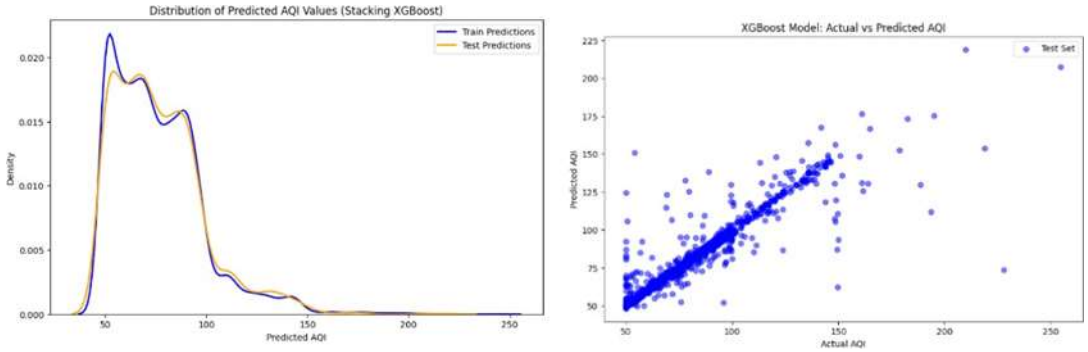

Fig. 5 (a): Density plot for RFR (b): Scatter plot for RFR

Density plot shows where predicted AQI values (orange) closely match the actual values (blue), indicating RFR's accuracy across all data points, including extremes. The scatter plot of actual vs. predicted AQI shows points clustered tightly along the 45-degree line, confirming high accuracy. The minimal spread across AQI values, even at higher levels, reflects RFR's strong predictive ability. This is consistent with the high R^2 score of 95% signifying excellent model performance.

3.4.3 XG Boosting

XG Boost (Extreme Gradient Boosting) is a powerful machine learning algorithm that excels in both regression and classification tasks. It builds models using a gradient boosting framework, where each new model focuses on correcting the errors of the previous ones, enhancing prediction accuracy. One of its key strengths is its ability to effectively handle missing data, making it suitable for real-world applications where datasets often have incomplete information. Additionally, XG Boost includes regularization techniques to reduce overfitting, ensuring that the model generalizes well to unseen data. The algorithm also supports parallel processing, significantly speeding up the training process compared to traditional methods. Its extensive

hyperparameter tuning options allow users to optimize model performance for specific datasets. As a result, XG Boost has become widely adopted in various fields, including finance, healthcare, and environmental monitoring, due to its effectiveness and efficiency in predictive analytics.

XG Boost was employed due to its high computational efficiency and superior predictive accuracy, particularly when working with tabular datasets. As a gradient boosting algorithm, XG Boost constructs decision trees sequentially, with each tree correcting the errors of its predecessors. This iterative process significantly improves AQI prediction accuracy. The model achieved an impressive **R^2 score of 97%**, indicating that it explained 97.04% of the variance in AQI values. Performance metrics were evaluated across training, validation, and test sets, confirming XG Boost's strong effectiveness compared to other models.

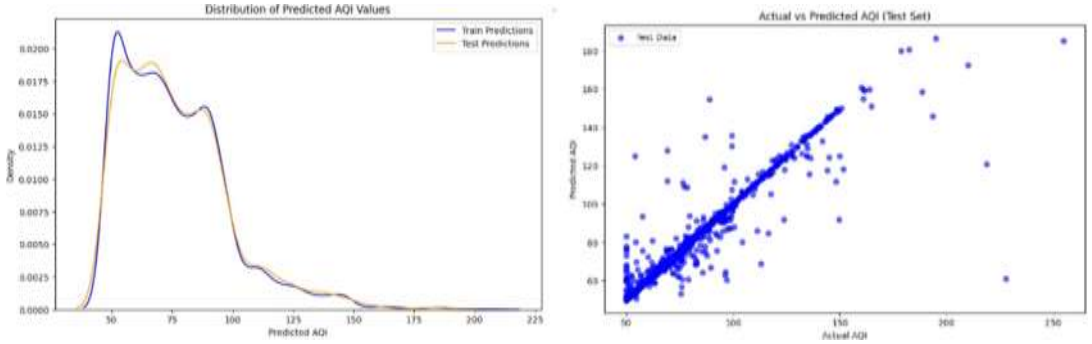

Fig. 6 (a): Density plot for XG Boost (b): Scatter plot for XG Boost

The left graph shows the density distribution of predicted AQI values using XG Boost, comparing the train and test predictions. The curves closely follow each other, indicating consistent model performance across both datasets. The right graph plots the actual versus predicted AQI values for the test set, illustrating a strong linear relationship as evidenced by the dense clustering along the diagonal, which signifies high accuracy in predictions.

3.5 ENSEMBLE MODELS

Ensemble learning techniques combine multiple predictive models to improve overall performance. In this study, **stacking** was employed as the primary ensemble method to leverage the strengths of different base learners for predicting Air Quality Index (AQI).

STACKING

Stacking is a method used to combine predictions from multiple models to improve accuracy. It works by using several base models to make predictions, which are then combined by a final model (called a meta-model) that produces the final prediction. The idea is to take advantage of the strengths of each base model while reducing individual weaknesses, resulting in better overall performance. Stacking usually involves different types of models, which allows it to capture a wider range of patterns in the data. This approach is especially effective in complex prediction tasks where a single model may not be sufficient.

1. **Stacking RFR+XG Boost**: The Stacking RFR + XG Boost approach is designed to improve predictive accuracy by combining the strengths of both the Random Forest Regressor (RFR) and the XGB Regressor. Each of these base models contributes unique capabilities: Random Forest excels at handling large datasets and reducing variance, while XG Boost offers high efficiency and performance through gradient boosting. In this setup, both models are trained separately on the data, and their predictions are then passed to a Ridge regression model, which acts as a meta-learner. The **Ridge model** learns to combine the outputs of RFR and XG Boost to make a more accurate final prediction.

By using this ensemble approach, the model benefits from both the diversity of the base learners and the stability of the meta-model. The result is a significant improvement in performance, achieving an **RMSE of 3.1950**, which indicates a low prediction error, and an **R^2 score of 98%,** showing that the model explains almost all the variance in the data. This makes the stacking model particularly effective in scenarios requiring high accuracy, such as air quality forecasting.

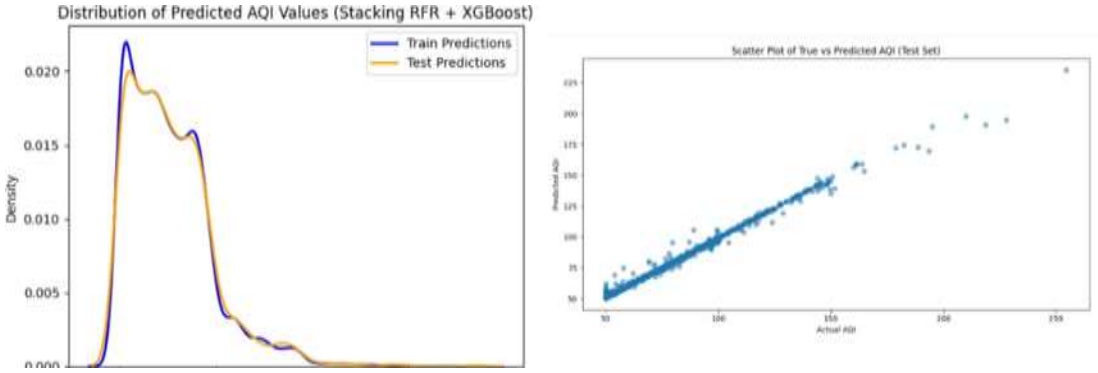

Fig. 7 (a) Density plot for Stacking RFR + XG Boost (b): Scatter plot for Stacking RFR + XG Boost

2. **Stacking SVR+RFR+XG Boost**: In this stacking model, Support Vector Regressor (SVR), Random Forest Regressor (RFR), and XGB Regressor are combined to leverage the strengths of each model. Each of these algorithms provides diverse predictions, which are then aggregated using a meta-model, often Ridge regression, to produce a final output. This layered approach improves the overall accuracy, as the meta-model learns how to best combine the base model predictions. The result is a lower RMSE (Root Mean Squared Error) compared to using the models individually and an **R^2 score of 84.2%,** indicating a high level of predictive accuracy.

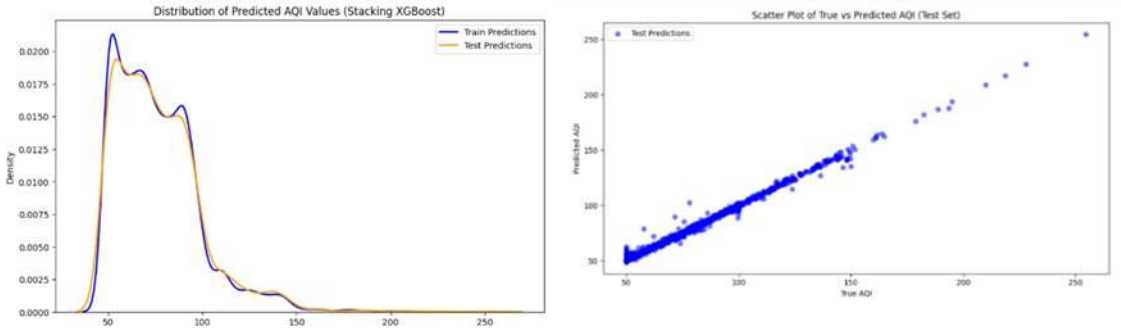

Fig. 8 (a): Density plot for Stacking SVR + RFR + XG Boost (b): Scatter plot for Stacking SVR + RFR + XG Boost

The ensemble methods, **Stacking (RFR + XG Boost)** and **Stacking SVR+RFR+XG Boost**, consistently outperform individual models, demonstrating the tightest alignment between actual and predicted values, which indicates superior predictive power. Among all models, **Stacking (RFR + XG Boost)** is the most reliable, closely followed by **Stacking SVR+RFR+XG Boost**.

These ensemble approaches effectively leverage the strengths of multiple algorithms, resulting in highly accurate AQI forecasts.

3.6 EVALUVATION METRICS

Evaluation metrics are quantitative measures used to assess the performance of machine learning models. They help determine how well a model's predictions align with actual outcomes. For regression tasks, common metrics include MAE, MSE, RMSE, and R^2, which evaluate prediction errors in different ways. The choice of metric depends on the specific goals, such as minimizing large errors or understanding overall variance explained.

1. **Mean Absolute Error (MAE)**

MAE measures the average magnitude of the absolute differences between predicted and actual values. It represents how far, on average, predictions deviate from actual values.

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

From Eq (2) Where where, n is the number of data points, yi is the actual value, y^i is the predicted value.

2. **Mean Squared Error (MSE)**

MSE calculates the average of the squared differences between predicted and actual values. It penalizes larger errors more heavily than smaller ones due to squaring the differences.

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

From Eq (3) Where where, n is the number of data points, yi is the actual value, y^i is the predicted value.

3. **Root Mean Squared Error (RMSE)**

RMSE is the square root of the MSE, bringing the error back to the same units as the predicted variable. It is more sensitive to large errors and provides a clearer interpretation of model performance.

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

From Eq (4) Where where, n is the number of data points, yi is the actual value, y^i is the predicted value.

4. **R-Squared (R^2)**

R^2 (the coefficient of determination) measures the proportion of variance in the dependent variable that is predictable from the independent variables. It assesses how well the model's predictions fit the actual data.

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

From Eq(5) where yi is the actual value, y^i is the predicted value, y¯ is the mean of the actual values, n is the number of data points.

3.7 AQI FORECASTING

For AQI forecasting, machine learning models like **SVR**, **Random Forest Regression**, and **XG Boost** were employed, with a stacked model combining **XG Boost** and **Random Forest** proving most effective. These models were trained on historical air quality data (2020–2023) from nine monitoring stations in Chennai.

To predict current and future AQI, data preprocessing involved managing missing values, calculating AQI from pollutant concentrations, and extracting relevant features. The models used historical data from the same month in previous years to predict current AQI, while the stacked models forecast future AQI based on historical trends for the same month in upcoming years.

Users can input a station name, month, and year, and the models predict the AQI, categorizing it into air quality levels (e.g., good, moderate, hazardous), with relevant health advice provided.

4. RESULTS AND DISCUSSION

4.1 INTERPRETATION

Initial evaluations of the models revealed significant variations in their performance, particularly concerning error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The Support Vector Regression (SVR) model exhibited the highest error rates, indicating its suboptimal performance for AQI prediction. Specifically, SVR achieved a low R^2 score of 45%, reflecting its limited ability to explain the variance in AQI values. This prompted the exploration of more advanced models to enhance predictive accuracy.

In this context, Random Forest Regression (RFR) and XG Boost were introduced. These models demonstrated substantial improvements, significantly reducing the error metrics. The Random Forest model achieved an R^2 score of 95.7%, while XG Boost outperformed with an R^2 of 97.2%, showing better predictive capability. Moreover, XG Boost exhibited lower MAE and RMSE values compared to other models, making it particularly effective in reducing prediction errors.

Table. 1 Evaluvation of Machine Learning Models

Metrics	SVR	RFR	XG Boosting	Stacking RFR+XGBoost	Stacking XGBoost
MAE	10.2129	0.8962	1.0584	1.0717	1.0521
MSE	282.5358	22.0337	14.4655	10.2078	13.2546
RMSE	16.8088	4.6940	3.8033	3.1950	3.6407
R^2	0.4538	0.9574	0.9720	0.9803	0.9744

To further refine the model performance, a stacking ensemble technique was employed, combining Random Forest and XG Boost. This ensemble, referred to as "Stacking RF + XG Boost," yielded the most accurate predictions, further reducing the RMSE to 3.1950 and achieving an R^2 score of 98%. This result highlights the robustness of ensemble models in air quality forecasting, as they significantly improved upon the performance of individual algorithms.

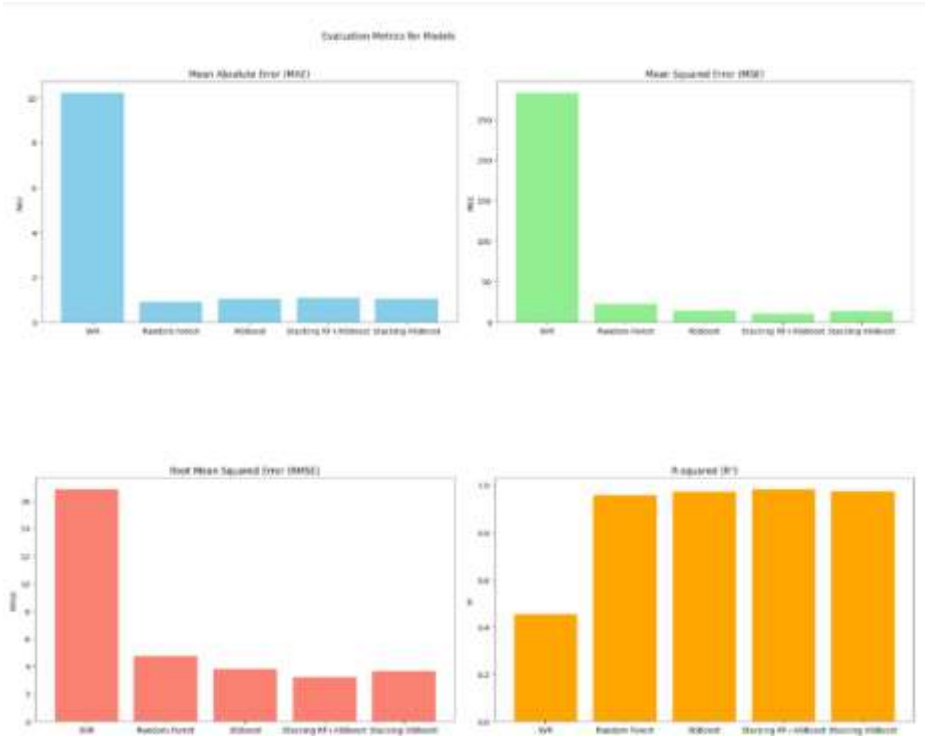

Fig. 9 Graphical Respresentation Evaluvation metrics of Machine Learning models

The graph illustrates the performance of several machine learning models in predicting Air Quality Index (AQI), evaluated using Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared (R^2). The models include Support Vector Regression (SVR), Random Forest Regression (RFR), XG Boost, and two ensemble methods combining RFR and XG Boost. Notably, the ensemble methods demonstrate superior performance across all metrics, highlighting their effectiveness in AQI forecasting.

4.2 USER INTERFACE

To enhance user interaction and demonstrate the practical application of our AQI prediction model, I developed a Graphical User Interface (GUI) using the Gradio library in Python.

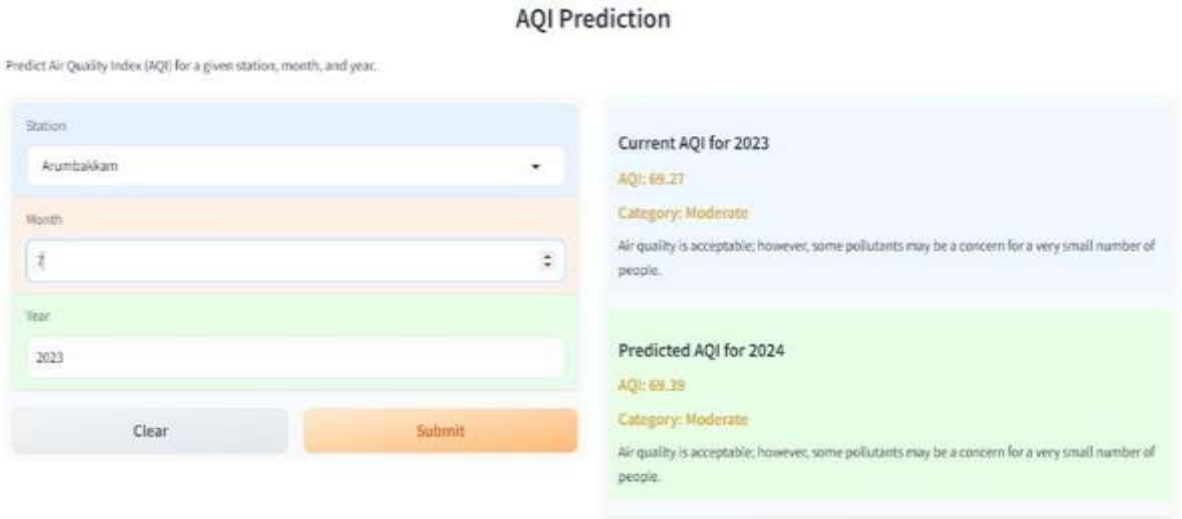

Fig. 10 User Interface for AQI Prediction

The GUI enables users to input key parameters, including the station name, month, and year, to obtain real-time AQI predictions. This interface leverages an advanced Ensemble machine learning model, combining Random Forest Regressor (RFR) and XG Boost, to provide accurate AQI estimates for the given year and a forecast for the following year, along with the corresponding air quality category. The Gradio-based GUI offers an intuitive and accessible platform, making it easier for users to interact with the system and gain valuable environmental insights.

5. CONCLUSION

Air quality has a direct impact on human life and society as a whole. Therefore, tackling air pollution requires collaborative efforts from the government, individuals, and organizations. The AQI index plays a crucial role in evaluating air quality and can guide the design of intelligent meteorological monitoring systems. In this research, a robust air quality prediction system was developed using machine learning techniques to accurately forecast AQI values in urban settings. By employing an ensemble stacking approach that combines Random Forest Regressor and XGB Regressor with Ridge as the meta-model, the system achieved improved prediction accuracy, with an impressive R^2 score of 98.03%, the highest among the models compared.

The system allows users to input station names, months, and years to receive accurate AQI forecasts, along with categorized health risk levels. The integration of pollutant data from monitoring stations across Chennai enhances prediction reliability, making it a valuable tool for managing urban air quality and guiding public health interventions. This scalable air quality monitoring solution is adaptable to other urban areas and can be extended to additional regions.

Future research could improve AQI prediction by incorporating deep learning models such as CNNs for spatial features and LSTMs or GRUs for temporal patterns. Hybrid models combining CNNs and LSTMs could capture both spatial and temporal air quality dynamics. Incorporating multi-modal data sources, including meteorological and traffic data, could enhance prediction accuracy. Transfer learning could reduce the need for large datasets when applying models to different regions. Real-time sensor data integration could enable dynamic public alerts and real-time AQI predictions, improving pollution control and public health protection.

REFERENCES

[1] https://airquality.cpcb.gov.in/.

[2] Mayuresh Mohan Londhe. “DM and ML Approach for Air Quality Index Prediction.” (2021).

[3] Shrddha Sagar, Dr. Deepali Viramani. “Air Quality Prediction using Machine Learning Algorithms – A Review.” In: IEEE (2020).

[4] Samayan Bhattacharya, Sk Shahnawaz. “Using Machine Learning to Predict Air Quality Index in New Delhi.” (2021).

[5] Natacha Soledad Represa, Alfonso Fern´andez Sarria, Andres Porta, Jes´us PalomarV´azquez. “Data Mining Paradigm in the Study of Air Quality.” (2019).

[6] Sachin Bhimrao Bhoite, Pooja Bhalgat, Sejal Pitale. “Air Quality Prediction Using Machine Learning Algorithm.” 2019.

[7] N. Srinivasa Gupta, Yashvi Mohta, Khyati Heda, Raahil Armaan, B. Valarmathi, G. Arulkumaran. “Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis.” 2023.

[8] K. Kumar, B. P. Pande. “Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities.” In: (2022).

[9] Avan Chowdary Gogineni, Vamsi Sri Naga Manikanta Murukonda. “Prediction of Air Quality Index Using Supervised Machine Learning.” (2022).

[10] To-Hieu Dao, Hoang Van Nhat, Hoang Quang Trung, Vu Hoang Dieu, Nguyen Thi Thu, Duc-Nghia Tran, Duc-Tan Tran. “Analysis and Prediction for Air Quality Using Various Machine Learning Models.” (2022).

[11] N. H. Van, P Van Thanh, D. N. Tran, DT. Tran. “A New Model of Air Quality Prediction Using Lightweight Machine Learning.” (2022).

[12] B. Raviteja, P. Tejaswini, U. Swetha Reddy. “Air Quality Prediction Using Machine Learning.” (2024).

[13] N. H. Van, P. Van Thanh, D. N. Tran, and D.-T. Tran, “A new model of air quality prediction using lightweight machine learning,” *International Journal of Environmental Science and Technology*, 2022.

[14] N. C. Minh, T. H. Dao, D. N. Tran, Q. H. Nguyen, T. T. Nguyen, and D. T. Tran, “Evaluation of Smartphone and Smartwatch Accelerometer Data in Activity Classification,” in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 2021, pp. 33–38.

[15] N. T. Thu, T.-h. Dao, B. Q. Bao, D.-n. Tran, P. V. Thanh, and D.-T. Tran, “RealTime Wearable-Device Based Activity Recognition Using Machine Learning Methods,” *International Journal of Computing and Digital Systems*, vol. 12, no. 1, pp. 321–333, 2022.

[16] J. K. Sethi and M. Mittal, “A new feature selection method based on machine learning technique for air quality dataset,” *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 697–705, 2019.

[17] H. Liu, Q. Li, D. Yu, and Y. Gu, “Air quality index and air pollutant concentration prediction based on machine learning algorithms,” *Applied Sciences (Switzerland)*, vol. 9, no. 19, 2019.

[18] M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L. Vanneschi, “A Machine Learning Approach to Predict Air Quality in California,” *Complexity*, vol. 2020, pp. 1–23, 2020.

[19] S. Zhang, “Nearest neighbor selection for iteratively kNN imputation,” *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, 2012.

[20] Y. Liu, Y. Wang, and J. Zhang, “New machine learning algorithm: Random forest,” *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence* and *Lecture Notes in Bioinformatics*), vol. 7473 LNCS, pp. 246– 252, 2012.