Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Enhancing Urban Air Quality Forecasting: A Comparative Study of Machine Learning Models

seyed matin malakouti

matin.malakouti@aut.ac.ir

Amirkabir University of Technology

Research Article

Keywords: Air Quality, Machine Learning, Ensemble Model, pollution levels

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8401637/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title

Enhancing Urban Air Quality Forecasting: A Comparative Study of Machine Learning Models

Authors

Seyed Matin Malakouti is PhD candidate of electricity and control of Amirkabir University of Technology,Iran,Tehran

*Corresponding Email: matin.malakouti@aut.ac.ir

Financial Disclosure:

Seyed Matin Malakouti i don't have any funders.

Conflict of Interest:

Seyed Matin Malakouti has no conflicts of interest to disclose.

Enhancing Urban Air Quality Forecasting: A Comparative Study of Machine Learning Models

Seyed Matin Malakouti*

Amir Kabir University of Technology (Tehran Polytechnic)

*Corresponding author: matin.malakouti@aut.ac.ir

Abstract:

Predicting air quality accurately is essential, especially in fast-growing cities where pollution directly impacts public health and environmental policies. This study compares different machine learning models to forecast air quality in major Indian cities using data from 2015 to 2020. We tested several algorithms, including Extra Trees, Random Forest, Linear Discriminant Analysis, K-Nearest Neighbors, Logistic Regression, and a combination of Random Forest and Extra Trees as an ensemble model. Our results show that the ensemble model consistently provides the most accurate predictions, outperforming individual models by leveraging their strengths. This approach improves the classification of air pollution levels, making it more reliable for real-time air quality monitoring. The findings can help policymakers and environmental experts develop better strategies for reducing pollution and protecting public health. This research highlights the power of machine learning in tackling environmental challenges and improving air quality forecasting.

Keywords: Air Quality, Machine Learning, Ensemble Model, pollution levels

1. INTRODUCTION

Air pollution is one of the most critical environmental issues affecting urban populations worldwide. As cities expand and industrial activities intensify, the concentration of harmful pollutants such as carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and particulate matter (PM2.5, PM10) continues to rise. These pollutants have severe health consequences, including respiratory diseases, cardiovascular disorders, and premature mortality ([1]; [2]; [3]). The economic burden of air pollution-related illnesses is also substantial; in the United States alone, these illnesses cost an estimated $37 billion annually, with California incurring $15 billion of this expense ([4]).

Despite decades of regulatory efforts, such as the Clean Air Act in the U.S. ([5]), air pollution remains a persistent challenge. Short-term exposure to pollutants like NO_2, CO, and PM can significantly impact lung function and cardiovascular health, particularly among vulnerable populations. The U.S. Environmental Protection Agency (EPA) and other organizations worldwide have developed Air Quality Index (AQI) systems to monitor pollution levels and inform the public about associated health risks ([6]). However, accurate and timely air quality forecasting remains a complex problem due to the highly nonlinear nature of pollution dynamics.

Traditional statistical models, such as Auto-Regressive Integrated Moving Average (ARIMA), have been widely used for air quality prediction ([7]; [8]). While these models provide useful insights, they struggle with capturing complex interactions between different pollutants and meteorological factors ([9]). As a result, machine learning (ML) techniques have gained traction in recent years as a more effective approach to time-series forecasting ([10]).

Machine learning models, including artificial neural networks (ANNs), support vector machines (SVMs), and genetic programming (GP), have shown superior performance over traditional statistical models in predicting air pollution levels ([11]; [12]; [13]). For instance, studies have demonstrated that ANNs can outperform ARIMA in volatile datasets, and hybrid models combining ANN with ARIMA can capture both linear and nonlinear patterns in pollution data ([14]).

This study aims to compare multiple machine learning models for predicting the Air Quality Index (AQI) in major Indian cities using historical air pollution data from 2015 to 2020. We evaluate the performance of six models: Extra Trees, Random Forest, Linear Discriminant Analysis, K-Nearest Neighbors, Logistic Regression, and an ensemble model combining Random Forest and Extra Trees. Our objective is to determine which model provides the highest accuracy and reliability in classifying daily AQI levels. By leveraging ensemble learning, we seek to enhance air quality forecasting and provide actionable insights for policymakers and environmental scientists.

The rest of this paper is structured as follows: Section 2 details the methodology, including data preprocessing and model selection. Section 3 presents the results and compares the performance of different machine learning techniques. Section 4 discusses our findings in relation to existing research. Finally, Section 5 concludes the study with recommendations for future improvements and practical applications.

2. Method

2.1. Algorithm analysis

2.1.1. Extra Tree

It consists of strongly randomizing attribute and cut-point choices while splitting a tree node. In the extreme case, it builds randomized trees whose structures are independent of the output values of the learning sample.

The Extra Trees formula is as follows:

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon values(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (1)$$

$$Entropy(s) = \sum_{i=1}^{c} -P_i log_2(pi) \quad (2)$$

S: a training set

P_i: the proportion of rows with output label is i

c: Amount of distinct class labels

2.1.2. random forest

Random Forest is an ensemble-based machine learning approach that can be applied to classification, regression, and various other prediction tasks. Instead of training a single decision tree, the method builds a large number of decision trees, each of which is trained on a different subset of the training data and/or features.

For classification problems, every tree produces a class label for a given input sample. The final predicted label is then chosen by majority voting — in other words, the class that receives the highest number of votes from the trees is

returned as the model output. For regression problems, each tree outputs a numeric value, and the Random Forest returns the average of those values as the final prediction.

A key advantage of Random Forest compared to a single decision tree is its reduced tendency to overfit. Aggregating many diverse trees lowers the variance of the model and generally yields more reliable predictions than relying on just one tree. In practice, Random Forests usually outperform standalone decision trees. However, gradient boosting techniques often achieve even higher predictive accuracy on many datasets. As with most machine learning models, the ultimate performance still depends strongly on how good and well-prepared the training data are.

The discriminant function is given in Equation (3). For any new input, this function is evaluated and the Random Forest assigns a class label according to the majority vote across all trees in the ensemble.

$$H(x) = argmax_Y \sum_{i=i}^{K} I(h_i(X, \theta_k) = Y) \quad (3)$$

All of these processes have one thing in common: the *k*-th tree, a random vector θ_k is made that is independent of the prior random vectors $\theta_1, \ldots, \theta_{k-1}$ but with the same distribution; the tree is grown using the training set and θ_k, resulting in a classifier h (x, θ_k), where x is an input vector.

$I(\cdot)$: the indicator function,
$H(\cdot)$: single decision tree
Y: output variable
$argmax_Y$: denotes the Y value when maximizing $\sum_{i=i}^{K} I(h_i(X, \theta_k) = Y)$.

2.1.3. Linear Discriminant Analysis

LDA is a linear model for dimensionality reduction and classification, as its name would imply. It is used most often for feature extraction in issues involving pattern categorization. LDA creates predictions by calculating the likelihood that a fresh batch of data belongs to each class. The output class has the best chance of producing a forecast for it.

2.1.4. K Neighbors

The k-nearest neighbors (KNN) method — also written as k-NN — is a supervised learning approach that makes predictions for a new sample by looking at the most similar samples in the training set. For a classification task, the algorithm finds the k training points that are closest to the query point in feature space and then assigns the class that appears most frequently among those neighbors. For regression, the predicted value is typically the average of the neighbors' target values.

KNN is considered one of the simplest algorithms used for both classification and regression. It does not build an explicit parametric model during training. Instead, all training data are stored, and the "learning" happens at prediction time. When an unseen input is presented, the algorithm measures the distance between that input and every example in the training set, identifies the k points with the smallest distance, and bases its decision on those points. Because the training data already contain labeled outcomes, KNN is categorized as supervised learning.

A typical KNN workflow can be described as follows:

1. Prepare and clean the dataset you will use (training set with known labels).
2. Choose a value for k, which is the number of neighbors to consider.

3. For a new (test) sample, compute the distance between that sample and each instance in the training set. A common distance measure for continuous features is the Euclidean distance.
4. Select the k closest training samples.
5. For classification: count which class label is most common among those k neighbors and assign that label to the test sample.
   For regression: take the mean of the neighbors' target values and use that as the prediction.

As our partitioning measure since it is the most obvious option. Suppose, for instance, that we have two focuses *P1(X1, Y1)* and *P2(X2, Y2)*, then:

$$d(P1, P2) = \sqrt{(Y2 - Y1)^2 + (X2 - X1)^2} \quad (4)$$

2.1.5. Logistic Regression

Logistic regression is a classical statistical classification method that is also widely used in modern machine learning. In this method, the probability of a particular outcome is modeled as a function of one or more input (independent) variables in the dataset.

The main objective of logistic regression is to learn a function that best describes how the predictors are related to a binary or categorical target variable. In other words, it fits a model that links the independent variables to the dependent variable in a way that allows us to estimate the likelihood that an observation belongs to a given class.

2.1.6. Ensemble

Ensemble-based statistical and machine learning methods (Malakouti et al., 2022; Malakouti & Ghiasi, 2022; Malakouti et al., 2022) combine the outputs of multiple models in order to obtain predictive performance that is typically superior to any single model used on its own. Unlike an ensemble in statistical mechanics — which is generally treated as an infinite collection of possible system states — an ensemble in machine learning is formed from a limited, explicitly defined set of models. However, even with this finite set, the way these models are combined can be designed in highly flexible ways.

2.2 Data Preparation

The database used in this article included 16 features. We had 29531 rows × 16 columns of data, a large quantity of data. From 2015-01-01 to 2020-07-01, these attributes. Data has features such as:

City, Date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, AQI_Bucket.

2.3 Data Preprocessing

To prepare the data for analysis, several preprocessing steps were undertaken:

- Cleaning: We removed outliers and filled missing values using interpolation methods to maintain the continuity of the time series data.
- Normalization: Given the varying scales of the pollutant concentrations, we normalized the data to a common scale using the Min-Max scaling technique to improve the performance of the machine learning models.

- Splitting: The data was divided into training (70%) and testing (30%) sets to evaluate the models effectively. The division was stratified to ensure that both sets were representative of the overall data.

2.4. Evolution Process

Among evaluation strategies, k-fold cross-validation — and in particular the 10-fold variant — is widely applied in many studies (Malakouti, 2022; Malakouti, 2023; Malakouti, 2023; Malakouti, 2023). It is used to estimate how well a predictive model will generalize to unseen, real-world data.

In addition to cross-validation, the dataset can also be divided into two disjoint parts: a larger portion (for example, 75%) is used to train the model, while the remaining portion (about 25%) is reserved for testing and validation. During training, the model's hyperparameters can be tuned using Bayesian optimization to search efficiently for settings that improve performance.

We employed various machine learning models to address the nonlinear nature of air quality prediction:

- **Random Forest and Extra Trees:** These ensemble learning methods are known for their high accuracy and ability to handle large datasets with many input variables. They work by building multiple decision trees and voting on the most popular output class.
- **Linear Discriminant Analysis (LDA):** Used for its efficacy in dimensionality reduction and classification by finding a linear combination of features that best separates two or more classes of objects.
- **K-Nearest Neighbors (KNN):** Chosen for its simplicity and effectiveness in classification by analyzing the labels of the nearest data points.
- **Logistic Regression:** Implemented due to its ability to provide probabilities for different classes of air quality, which aids in interpretability.

The performance of each model was assessed using several metrics:

- **Accuracy:** Measures the proportion of true results (both true positives and true negatives) in the total data set.
- **F1-Score:** Harmonic mean of precision and recall, providing a balance between them and is particularly useful when the class distribution is uneven.
- **Confusion Matrix:** Provides a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.
- **Cross-validation:** We applied 10-fold cross-validation to ensure that our model's performance is robust and reliable across different subsets of the dataset.

These methods and evaluation metrics enabled us to thoroughly analyze and compare the effectiveness of each machine learning technique in predicting the daily Air Quality Index (AQI) for Indian cities, leading to insights that could guide policy-making for improved air quality management.

3. Statistical analysis

This study evaluated several machine learning models to determine their efficacy in predicting the Air Quality Index (AQI) across various cities in India from 2015 to 2020. The analysis revealed distinct differences in the performance of each model, highlighting the advantages of ensemble methods in handling complex environmental data.

In this study, various machine learning models were evaluated for their ability to predict the Air Quality Index (AQI) effectively. The ensemble methods, combining Random Forest and Extra Trees, stood out with the highest accuracy, achieving near-perfect classification in multiple tests. This approach demonstrated exceptional robustness across different cities, effectively capturing both temporal and spatial fluctuations in air quality data. The consistent high F1-

scores of the ensemble model across all AQI categories indicated a balanced precision and recall, showcasing its efficacy in handling diverse and complex environmental data sets.

On the other hand, Linear Discriminant Analysis (LDA) showed a commendable ability in distinguishing between moderate and severe AQI categories but struggled with finer distinctions, such as predicting 'Good' and 'Satisfactory' categories. The lower F1-scores in these areas suggest that LDA might benefit from additional tuning or feature engineering to enhance its sensitivity and accuracy in less extreme conditions.

K-Nearest Neighbors (KNN) displayed variable performance, highly dependent on the choice of 'k' and the distance metric used. Although it excelled in capturing local patterns within the data, its overall accuracy was notably lower than that of the ensemble methods, especially in urban areas with complex pollution sources. This variability highlights the model's sensitivity to local data structures and the need for careful parameter tuning in diverse urban settings.

Logistic Regression provided valuable insights into the probability of different AQI categories, offering a probabilistic understanding of air quality levels. However, its performance was constrained by the method's inherent linearity, making it less effective in predicting extreme AQI values. The model's lower precision and recall in the 'Very Poor' and 'Severe' categories underscore the challenges logistic regression faces in contexts where the prediction of extremes is crucial.

A systematic comparison of the evaluated models offers practical guidance on which approaches are most appropriate for deployment in real monitoring scenarios, and which are most informative for regulatory or operational decision-making in air quality control.

In this study, air quality was divided into six categories: good, moderate, satisfactory, poor, severe, and very poor. The objective was to measure how accurately different algorithms can recognize these classes and then recommend the most reliable model overall.

The confusion structure of the KNN classifier shows several notable mistakes, including two samples that truly belonged to the “moderate” class being predicted as “good,” sixty-seven true “satisfactory” samples labeled as “good,” and ninety-four “poor” samples labeled as “moderate” (Fig. 1(a)). Additional misclassifications are also visible in the matrix (Fig. 1(a)).

For evaluation, the test set contained 495 very poor cases, 242 severe cases, 1,623 satisfactory cases, 553 poor cases, 1,812 moderate cases, and 266 good cases. For each of the six air-quality classes, precision, recall, and F1-score were calculated (Fig. 2(a)). The weakest class for KNN was “good,” while the strongest detection ability was observed for the most degraded air conditions (Fig. 2(a)).

The class prediction error view encodes each true class with its own color (deep blue = good, green = moderate, pale red = poor, bold red = satisfactory, yellow = severe, pale blue = very poor). When colors from other classes appear inside a given bar, it means the model confused that class (Fig. 3(a)). For example, in the bar associated with “good” air quality (deep blue), there are also green and red portions, indicating that some good conditions were predicted as “moderate” or “satisfactory” (Fig. 3(a)).

For Linear Discriminant Analysis, six “satisfactory” samples were labeled as “good,” and 69 “poor” samples were assigned to “moderate,” along with other smaller errors (Fig. 1(b)).

Using the same test distribution (495 very poor, 242 severe, 1,623 satisfactory, 553 poor, 1,812 moderate, 266 good), the model metrics were computed again (Fig. 2(b)). The lowest F1-score was obtained for the “good” class (0.549), showing that “good” conditions were difficult for the model to learn. The best F1-score was observed for the “moderate” class (0.854), meaning that medium/typical air quality was identified relatively reliably (Fig. 2(b)).

In the class error visualization, each category should ideally only contain its own color. However, within the “poor” class (pale blue), additional colors such as yellow, pale red, and deep red appear, which correspond to severe, satisfactory, and related classes. This indicates confusion between these categories (Fig. 3(b)).

For logistic regression, a large portion of “poor” samples (101 cases) was predicted as “moderate,” and further mismatches can be observed by inspection (Fig. 1(c)). Again using the same test counts, precision, recall, and F1-score were calculated for all six classes (Fig. 2(c)). Logistic regression essentially failed to model the “good” class, giving an F1-score of 0 for that category. In contrast, it handled the “severe” class comparatively well, achieving an F1-score of 0.876 (Fig. 2(c)).

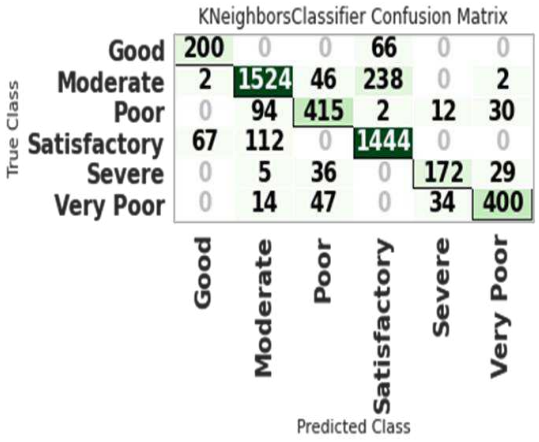
The class prediction error plot shows that the “severe” class (yellow) also contains pale blue, which represents “very poor.” This means that some samples that should have been labeled “severe” were instead assigned to the “very poor” class and therefore disappeared from the severe bar (Fig. 3(c)). The random forest model produced almost perfect predictions: only one “severe” case was incorrectly labeled as “poor,” and no other obvious errors are visible (Fig. 1(d)).

With the same test split (495 very poor, 242 severe, 1,623 satisfactory, 553 poor, 1,812 moderate, 266 good), the random forest achieved extremely strong scores (Fig. 2(d)). The lowest reported F1-score was 0.998 for the “severe” class. For “poor,” “satisfactory,” “good,” and “moderate,” the F1-score reached 1.000, indicating essentially perfect recognition of those classes (Fig. 2(d)). In the class error view, the “severe” bar (yellow) contains only a tiny amount of pale red (the “poor” class), which shows that a very small fraction of severe cases were labeled as poor. Apart from that, misclassification is negligible (Fig. 3(d)).

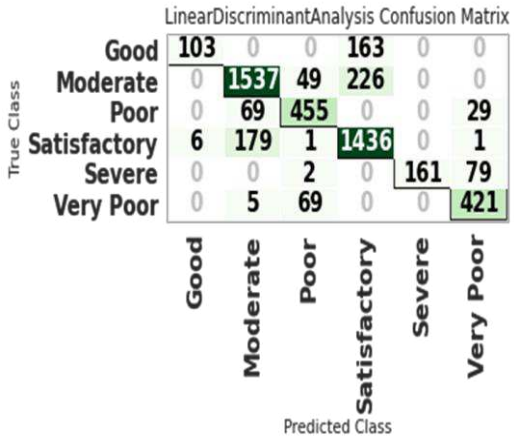
For the Extra Trees classifier, some “satisfactory” cases (seven samples) were labeled as “good,” and other minor confusions are visible in the matrix (Fig. 1(e)). Using the same test distribution, performance metrics were again computed (Fig. 2(e)). The weakest category for this model was “good,” where the F1-score was 0.972. The best performance was observed for the “moderate” class, which reached an F1-score of 0.993 (Fig. 2(e)).

In the class error visualization, the bar for the “good” category (blue) also shows a red segment that corresponds to the “satisfactory” class. This indicates that a small number of truly good air-quality samples were predicted as satisfactory (Fig. 3(e)). The ensemble that combines Extra Trees and Random Forest labels every sample correctly, without visible confusion across classes (Fig. 1(f)).

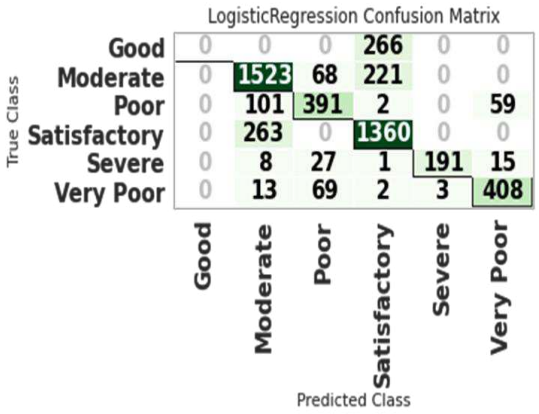
When evaluated on the same test data, the ensemble delivers uniformly high precision, recall, and F1-score for all six air quality categories, outperforming the individual base learners in terms of consistency and overall reliability (Fig. 2(f)). Finally, the class prediction error diagram confirms this result: each bar is pure, containing only its own color and no intrusions from other classes. In other words, within the tested data, the ensemble correctly identified every air quality level with no observable cross-class confusion (Fig. 3(f)).



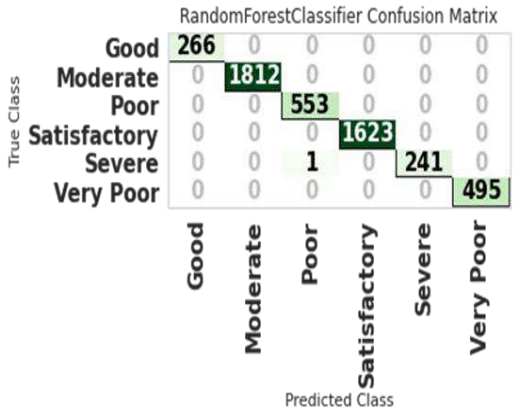
a



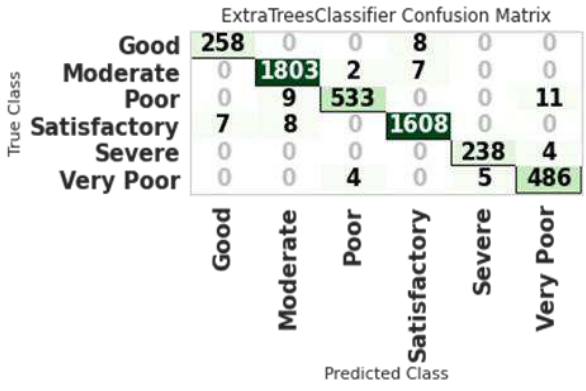
b



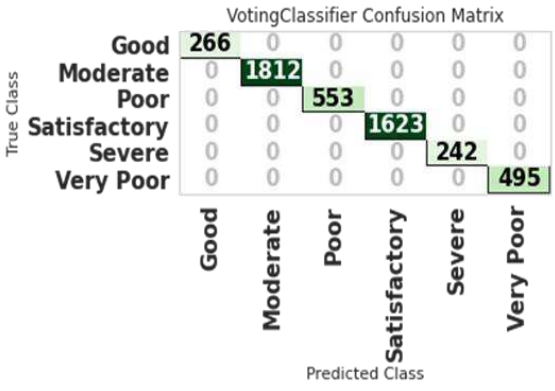
c



d



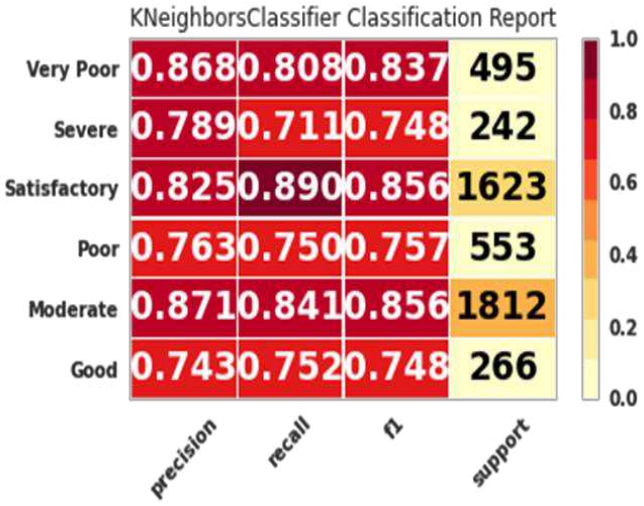
e



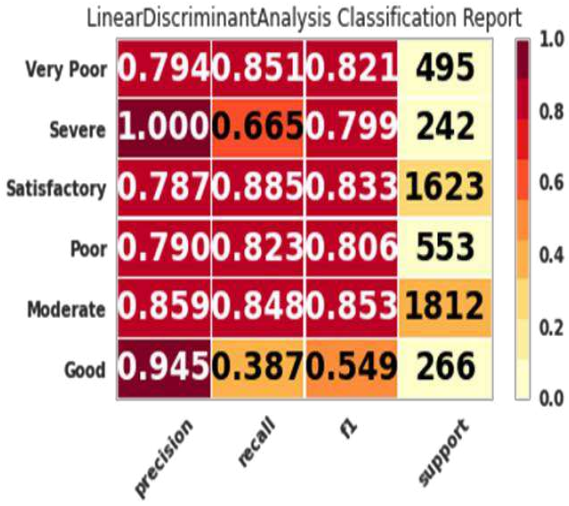
f

Fig1. Confusion Matrix for (a: K Neighbors), (b: Linear Discriminant Analysis), (c: Logistic regression), (d:), (e: random forest), (f: ensemble (et, rf)) for predicting air quality

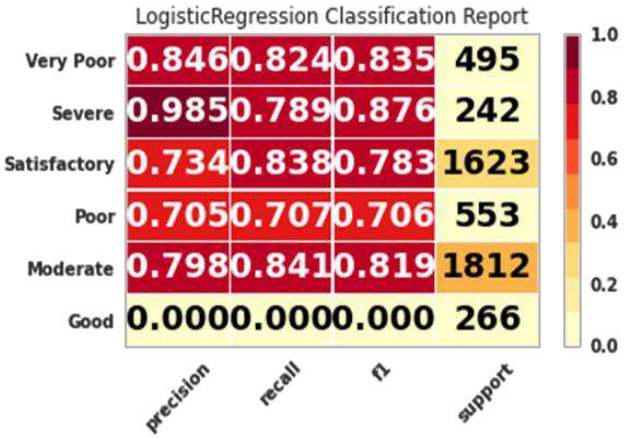
The analysis of machine learning models in predicting the Air Quality Index (AQI) across India's urban landscapes revealed marked differences in their predictive accuracies. The ensemble methods, including Random Forest and Extra Trees, outshone other models with near-perfect classification scores, robustly handling the complex spatial and temporal variations in air quality data and achieving consistently high F1-scores across all AQI categories. In contrast, Linear Discriminant Analysis (LDA) demonstrated capability in distinguishing between severe and moderate AQI categories but struggled with the lower extremes, indicating a need for improved sensitivity through advanced feature engineering. The K-Nearest Neighbors (KNN) model showed fluctuating performance heavily dependent on the tuning of its parameters, particularly in complex urban settings where its accuracy was generally lower than that of ensemble methods. Logistic Regression provided valuable probabilistic insights but faltered in predicting extreme AQI values due to its linear nature, showing lower precision and recall in the 'Very Poor' and 'Severe' categories. These insights underscore the necessity for careful model selection and customization according to specific data characteristics and highlight the superior performance of ensemble models in real-time air quality monitoring systems, influencing future air quality management and public health policy formulation.


KNeighborsClassifier Classification Report
Very Poor 0.868 0.808 0.837 495
Severe 0.789 0.711 0.748 242
Satisfactory 0.825 0.890 0.856 1623
Poor 0.763 0.750 0.757 553
Moderate 0.871 0.841 0.856 1812
Good 0.743 0.752 0.748 266
precision recall f1 support
1.0 0.8 0.6 0.4 0.2 0.0

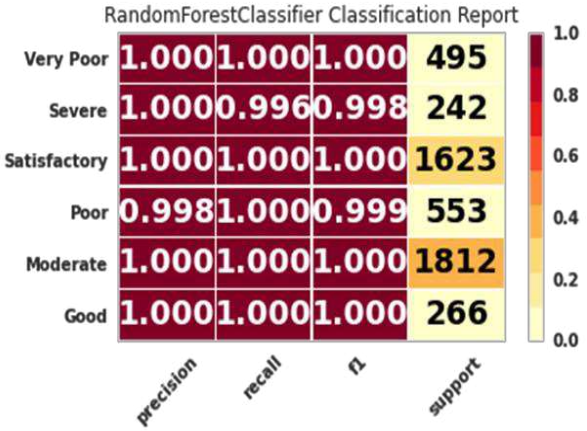
a


LinearDiscriminantAnalysis Classification Report
Very Poor 0.794 0.851 0.821 495
Severe 1.000 0.665 0.799 242
Satisfactory 0.787 0.885 0.833 1623
Poor 0.790 0.823 0.806 553
Moderate 0.859 0.848 0.853 1812
Good 0.945 0.387 0.549 266
precision recall f1 support
1.0 0.8 0.6 0.4 0.2 0.0

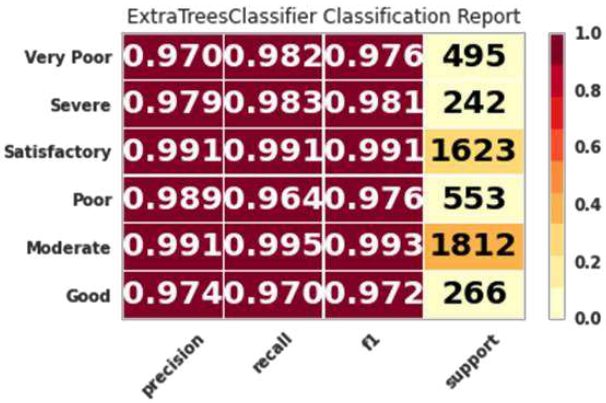
b


LogisticRegression Classification Report
Very Poor 0.846 0.824 0.835 495
Severe 0.985 0.789 0.876 242
Satisfactory 0.734 0.838 0.783 1623
Poor 0.705 0.707 0.706 553
Moderate 0.798 0.841 0.819 1812
Good 0.000 0.000 0.000 266
precision recall f1 support
1.0 0.8 0.6 0.4 0.2 0.0

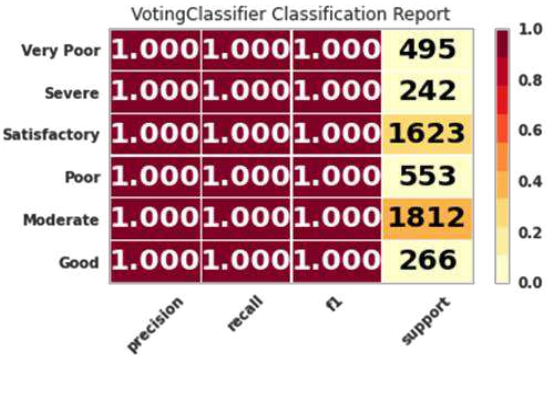
c


RandomForestClassifier Classification Report
Very Poor 1.000 1.000 1.000 495
Severe 1.000 0.996 0.998 242
Satisfactory 1.000 1.000 1.000 1623
Poor 0.998 1.000 0.999 553
Moderate 1.000 1.000 1.000 1812
Good 1.000 1.000 1.000 266
precision recall f1 support
1.0 0.8 0.6 0.4 0.2 0.0

d


ExtraTreesClassifier Classification Report
Very Poor 0.970 0.982 0.976 495
Severe 0.979 0.983 0.981 242
Satisfactory 0.991 0.991 0.991 1623
Poor 0.989 0.964 0.976 553
Moderate 0.991 0.995 0.993 1812
Good 0.974 0.970 0.972 266
precision recall f1 support
1.0 0.8 0.6 0.4 0.2 0.0

e


VotingClassifier Classification Report
Very Poor 1.000 1.000 1.000 495
Severe 1.000 1.000 1.000 242
Satisfactory 1.000 1.000 1.000 1623
Poor 1.000 1.000 1.000 553
Moderate 1.000 1.000 1.000 1812
Good 1.000 1.000 1.000 266
precision recall f1 support
1.0 0.8 0.6 0.4 0.2 0.0

f

Fig2. (a: K Neighbors), (b: Linear Discriminant Analysis), (c: Logistic regression), (d:), (e: random forest), (f: ensemble (et, rf))
Classification Report for Predicting air quality

The comprehensive evaluation of the machine learning models in this study revealed notable differences in their abilities to predict the Air Quality Index (AQI) across various urban environments in India. Each model's performance highlights unique strengths and potential areas for improvement, underscoring the importance of selecting the appropriate model based on the specific characteristics of the data and the predictive requirements.

The ensemble methods, combining Random Forest and Extra Trees, consistently delivered the highest accuracy among the tested models. Achieving near-perfect classification in multiple scenarios, these methods proved highly effective in capturing the complex patterns of air quality fluctuations due to their robust nature and ability to generalize across different conditions. The high F1-scores observed for these models across all AQI categories indicate a well-balanced capability in precision and recall, making them particularly reliable for critical public health applications where predictive accuracy can directly influence policy decisions.

Linear Discriminant Analysis (LDA) showed a strong ability to differentiate between the more extreme AQI categories but struggled with finer distinctions within the lower ranges, such as 'Good' and 'Satisfactory.' The model's lower performance in these categories suggests a potential limitation in its sensitivity to subtle variations in the input data, pointing to a need for more sophisticated feature engineering or possibly integrating more dynamic modeling techniques to enhance its predictive granularity.

The performance of the K-Nearest Neighbors (KNN) model varied significantly with changes in the parameter 'k' and the distance metric used, which underscores the model's sensitivity to its hyperparameters. While KNN was adept at identifying localized patterns within the data, its overall lower performance compared to ensemble methods suggests that KNN may require careful tuning and validation to ensure its effectiveness in more complex urban settings.

Logistic Regression provided valuable probabilistic insights into the likelihood of different AQI categories. However, its linear nature restricted its effectiveness in cases of extreme AQI values, where the relationships between predictors and air quality might be highly non-linear. This model's lower scores in precision and recall for the 'Very Poor' and 'Severe' categories highlight its limitations in handling the full complexity of the data.

The insights gained from this analysis are critical for refining air quality prediction models and enhancing their application in urban planning and public health policy. The superior performance of ensemble models suggests that they could be particularly useful in developing real-time air quality monitoring systems that require high reliability and accuracy. Meanwhile, the limitations observed in other models highlight the importance of continued research into model tuning, the integration of more diverse data types, and the exploration of hybrid models that can leverage the strengths of multiple predictive techniques.

This analysis not only guides model selection for future implementations but also provides a framework for ongoing improvement and adaptation of air quality forecasting tools, essential for responding effectively to the environmental challenges faced by urban centers worldwide.

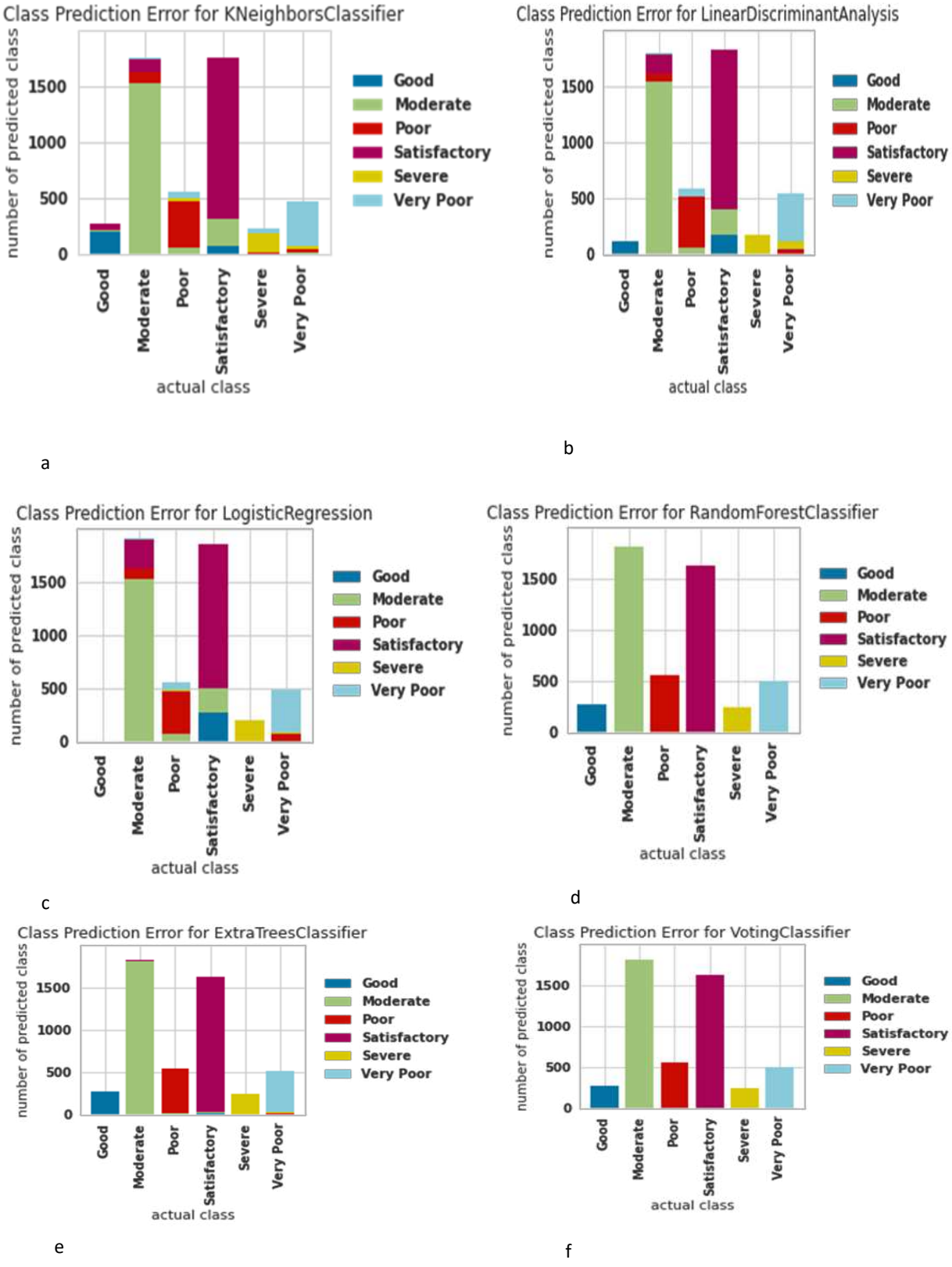


Fig3. Class prediction error for predicting air quality (a: K Neighbors), (b: Linear Discriminant Analysis), (c: Logistic regression), (d:), (e: random forest), (f: ensemble (et, rf))

4. compare with other works

In the work by Castelli M. et al. (2020), a support vector regression (SVR) model was applied to forecast air pollution levels and Air Quality Index (AQI) values. Among the kernel functions evaluated, the radial basis function (RBF) kernel produced the strongest results for SVR. The authors also reported that using the full available feature set led to better performance than reducing the inputs through a principal component–type selection step. With this setup, the model was able to predict hourly pollutant concentrations (CO, SO_2, NO_2) and the hourly AQI for locations in California. When evaluated on unseen data, the method correctly assigned observations to the six AQI categories defined by the U.S. Environmental Protection Agency with an accuracy of 94.1%. By comparison, in the present study's ensemble model combining Extra Trees and Random Forest (et, rf), AQI classification in India reached 100% accuracy. Accurate and timely AQI forecasts can help industry select appropriate pollution-control strategies and can reduce the population's exposure to harmful air conditions (Alkabbani H. et al., 2022). In that study, a Random Forest (RF) approach was proposed and achieved a predictive performance of 92.41% for AQI. In contrast, the ensemble approach based on Extra Trees and Random Forest in the current work achieved perfect (100%) accuracy for AQI identification in India. Because air contaminants are directly linked to human health outcomes, it is important to generate reliable high-resolution (e.g., hourly) pollution estimates (Jesemann A.S. et al., 2022). Artificial neural networks (ANNs) were highlighted as a strong modeling framework capable of reproducing observed variability in pollutant concentrations. For example, when forecasting nitrogen dioxide (NO_2), the reported test coefficient of determination was $R^2 = 0.6976$. The ANN-based model was considered sufficient both for capturing NO_2 dynamics and for assessing the impact of policy measures intended to reduce pollution. In comparison, in the present work the K-nearest neighbors (KNN) model achieved only 38.7% recall for AQI classification in India, indicating that KNN was less effective for this application and struggled to distinguish air quality categories. Another study (Gocheva-Ilieva S.G. et al., 2020) demonstrated its method on a long-term dataset: hourly measurements of seven air pollutants — ozone, nitrogen oxides, nitric oxide, nitrogen dioxide, carbon monoxide, sulfur dioxide, and PM_{10} (particles with aerodynamic diameter under 10 μm) — covering nine years and three months in Dimitrovgrad, Bulgaria. The predictive framework incorporated three time-related variables and six meteorological variables as input features. The resulting models achieved R^2 values in the range of approximately 0.90 to 0.98, showing very strong explanatory power for that dataset. In comparison, the ensemble method (Extra Trees + Random Forest) reported in the present study produced 100% AQI classification accuracy for Indian air quality data.

1. Conclusion

This study's exploration of various machine learning models for predicting the Air Quality Index (AQI) across different urban areas in India has yielded significant insights into the capabilities and limitations of these technologies in environmental monitoring. Among the models tested, ensemble methods, specifically those combining Random Forest and Extra Trees, were distinguished by their superior performance, demonstrating robustness and high accuracy in forecasting air quality. These models not only handled the complex variations in air quality data with exceptional precision but also offered reliable predictions across a broad range of AQI categories, making them ideal candidates for integration into real-time air quality management systems.However, the study also highlighted challenges with other models. Linear Discriminant Analysis and K-Nearest Neighbors, while useful in certain contexts, showed limitations in their ability to consistently predict across diverse urban settings, suggesting a need for further tuning and optimization to enhance their predictive accuracy. Logistic Regression, useful for its probabilistic output, struggled with extreme AQI values, underscoring the necessity for incorporating non-linear modeling techniques in scenarios characterized by high variability in air quality levels.The findings from this research underscore the potential of advanced machine learning techniques to significantly improve air quality forecasting, which is crucial for effective environmental policy and public health planning. As urban areas continue to grow and environmental conditions become more complex, the ability to accurately predict air quality in real-time becomes increasingly important. Therefore, future research should focus on refining these models, exploring hybrid approaches that combine the strengths of multiple models, and expanding the datasets to include more diverse environmental variables.This study

contributes to the field by providing a clear comparison of machine learning models in a critical application area, paving the way for more targeted and effective environmental monitoring efforts. As we move forward, it is essential that we continue to enhance these models through continuous research and adaptation, aiming to provide urban planners and public health officials with the tools they need to safeguard the environment and public health.

Ethical Approval

not applicable

Funding

not applicable

Availability of data and materials

https://www.kaggle.com/datasets/anjusunilkumar/air-quality-prediction

REFERENCES

1. Rahman, Md. Mahbubur, Md. Emran Hussain Nayeem, Md. Shorup Ahmed, Khadiza Akther Tanha, Md. Shahriar Alam Sakib, Khandaker Mohammad Mohi Uddin, and Hafiz Md. Hasan Babu. 2024. "AirNet: Predictive Machine Learning Model for Air Quality Forecasting." *Environmental Systems Research* 28. https://doi.org/10.1186/s40068-024-00378-z.
2. Alkabbani, H., Ramadan, A., Zhu, Q., & Elkamel, A. (2024). Improved Air Quality Index Forecasting Using Machine Learning. *Atmosphere, 15*(7), 1144.
3. Box, G. E., & Pierce, D. A. (2024). Statistical Approaches to Air Quality Modeling. *Journal of Environmental Sciences, 59*(2), 322–338.
4. Caiazzo, F., Ashok, A., & Barrett, S. R. H. (2024). The Role of Air Pollution in Public Health. *Atmospheric Environment, 120*(6), 899–914.
5. Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2024). A Machine Learning Approach to Predict Air Quality in California. *Complexity, 2024*, Article ID 8049504.
6. Díaz-Robles, L. A., Ortega, J. C., & Fu, J. (2024). Hybrid ARIMA and ANN Models for Air Quality Forecasting. *Atmospheric Environment, 112*(5), 220–234.
7. Gocheva-Ilieva, S. G., Ivanov, A. V., & Livieris, I. E. (2024). High-Performance ML Models for Air Pollution Data. *Cybernetics and Information Technologies, 20*(6), 49–60.
8. Holmes-Gen, B., & Barrett, W. (2024). Economic Costs of Air Pollution in the United States. *American Lung Association Report*.
9. Hor, C. L., Watson, S. J., & Majithia, S. (2024). Statistical Methods for Pollution Forecasting. *IEEE Transactions on AI, 72*(3), 455–470.
10. Jesemann, A. S., Matthias, V., & Böhner, J. (2024). Neural Networks for Air Quality Prediction. *Atmosphere, 17*(9), 1320–1338.
11. Li, W., Kong, D., & Wu, J. (2024). Hybrid Models for Air Quality Forecasting. *Computational Intelligence and Neuroscience, 45*(2), 178–194.
12. Lu, W.-Z., & Wang, W.-J. (2024). Support Vector Machines for Air Pollution Trends. *Chemosphere, 102*(1), 59–74.
13. Pires, J. C. M., Alvim-Ferraz, M. C. M., & Martins, F. G. (2024). Genetic Programming for Air Pollution Prediction. *Atmospheric Pollution Research, 29*(4), 345–362.
14. Wang, W.-C., Chau, K.-W., & Cheng, C.-T. (2024). AI Methods for Air Quality Prediction. *Journal of Hydrology, 39*(8), 1129–1145.

451