Enhancing Outlier Detection in Air Quality Index Data Using a Stacked Machine Learning Model

Abdoul Aziz Diallo[1]*
abdoul.aziz@students.jkuat.ac.ke

Lawrence Nderu[2]
lnderu@jkuat.ac.ke

Bonface Miya Malenje[3]
bmalenje@jkuat.ac.ke

Gideon Mutie Kikuvi[4]
kikuvi@jkuat.ac.ke

[1] Department of Mathematics (Data Sciences), Pan African University, Institute of Basic Science, Technology and Innovation (PAUSTI), Kenya
[2] Department of Sciences Technology, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya
[3] Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya
[4] Department of Environmental Health and Disease Control, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya

Abstract

Air quality is an important part of environmental health, having serious consequences for human health and well-being. The Air Quality Index (AQI) is a frequently used metric for assessing air quality in various areas and at different times. However, AQI data, like many other types of environmental data, can contain outliers data points that deviate significantly from other observations, indicating exceptionally good or poor air quality, a critical step in identifying and understanding extreme pollution episodes that can have serious environmental and public health consequences. These outliers can be caused by a variety of variables, including measurement mistakes, odd meteorological circumstances, and pollution occurrences. While outliers can occasionally give useful information about these unusual conditions, they can also skew studies and models if they are not adequately accounted for. This paper describes a hybrid method for detecting outliers in data, AQI data are used in this study. The model uses a stacked machine learning model that incorporates K-means clustering, Random Forest (RF), and Gradient Boosting Classifier (GBC). K-means is used for initial categorization, followed by RF model training, and ultimately, the RF output is used as input for the GBC to generate the final classification. The performance of this stacked machine learning model is examined and compared to single models using the Accuracy measure. The findings show that the suggested technique is efficient, with an accuracy of 0.99, showing its potential for effective outlier detection in data.

KEYWORDS: outlier detection, k-means, random forest, gradient boosting classifier, air pollution, data mining

1 Introduction

Air quality is a major factor in determining one's health and quality of life. Monitoring and understanding air quality has become increasingly vital as global urbanization and industrialization continue to rise. The AQI is a critical instrument in this attempt, offering a standardized, understandable assessment of air quality that policymakers, researchers, and the general public can all utilize [10]. The AQI is a composite metric that reflects the quantities of numerous pollutants in the air, such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), carbon monoxide (CO), and ozone (O3). Each of these pollutants can have various health effects, and their levels can fluctuate dramatically depending on factors such as weather, time of day, and human activity. As a result, AQI data can display significant variety and complexity (Table 3) [5, 9]. One key problem in analyzing AQI data is the occurrence of outliers, or data points that differ dramatically from the rest of the data. Outliers might occur as a result of measurement mistakes, odd meteorological circumstances, uncommon pollution incidents, or other abnormalities. While outliers can occasionally give significant insights into unusual occurrences or data-collecting concerns, they can also affect statistical studies and prediction models if not handled appropriately [6]. In this work, we offer a powerful strategy for detecting outliers in data by combining machine learning and clustering approaches. A layered model including K-means clustering, RF, and GBC is used in our technique. We first categorize the data using K-means, then train the RF model on these categories, and lastly utilize the output of the RF model as input to the GBC to make the final outlier classifications. We compare our stacked model's performance to single models such as Decision Trees, Support Vector Machines, K-Nearest Neighbors, and Naive Bayes using different metrics such as precision, recall, accuracy, and F1-score. Our findings illustrate the efficacy of this strategy, with good scores across all measures. This study adds to continuing efforts to enhance the accuracy and reliability of air quality evaluations and projections, which are critical for influencing public health treatments and environmental policy.

2 Literature Review

This study introduces a method for detecting outliers in urban NO2 concentrations using low-cost air quality sensors, based on spatiotemporal categorization into 16 groups. Outliers are identified using mean and standard deviation within these groups, with the method detecting 0.1-0.5% of outliers in a sensor network in Eindhoven. The research concludes that this approach effectively identifies outliers while maintaining the spatio-temporal variability of air pollution [19]. The study proposes a three-module method for outlier detection in indoor air quality data, using a long short-term memory auto-encoder and a vector machine detector to form two distinct models. These models are then unified using an ensemble-based decision rule in the third module, with laboratory tests validating the framework's consistent effectiveness in identifying outliers across various industrial scenarios [14, 16]. The study introduces FForest, a novel anomaly detection method that combines Isolation Forest with fuzzy set theory, focusing on isolating anomalies for efficient computation. It enhances stability by using quartile-based identification of potential anomalies instead of random sampling. Experimental results on seven real-world datasets show FForest's superior performance in anomaly detection compared to four baseline methods [17, 27].

This study introduces an automated outlier detection method for air quality data, based on discrepancies between observed and estimated pollutant concentrations. Applied to data from China's National Environmental Monitoring Network, it effectively identified outliers in six pollutants, revealing a trend of decreasing outliers over time. The method also highlighted the significant impact of outliers on annual mean concentrations of PM2.5 at multiple sites [3, 23]. The article emphasizes the importance of reliable data collection from surface sensors for managing air pollution and highlights the challenge of outliers in the data. Current outlier detection methods are noted to be inaccurate and susceptible to false positives. The authors propose a two-step strategy to discern

between real and false outliers, reducing false positives and enhancing the accuracy of air pollution forecasting, thereby facilitating effective solutions to this global problem [12, 29]. The study develops an automated outlier detection system for air quality networks, accounting for issues like instrument problems and harsh conditions. Applied to six pollutants across 1436 sites in China, the system detected outliers in 0.65%-5.68% of measurements, significantly influencing yearly mean PM2.5 concentrations at 66 locations [7, 23].

Findings from the current introduces a new hybrid air pollution forecasting model that enhances accuracy by integrating an outlier identification and correction technique with a heuristic intelligent optimization algorithm. The model outperforms existing ones, providing reliable forecasts and aiding in the development of effective strategies to reduce air pollutant emissions [15, 21]. The essay addresses the challenge of outliers in spatiotemporal data streams from geographically dispersed sensor networks, which can distort future analyses. The study proposes two novel IPCA-based outlier detection methods, compares them with existing techniques, and provides insights into IPCA's applicability for real-time applications such as image analysis, pattern recognition, and credit card fraud detection [4, 28]. This paper focuses on the air quality index (AQI), a numerical measure affected by human activities, industrial operations, and weather conditions, and uses regression models to estimate AQI based on data from a monitoring station in Chennai, India. The study investigates the significance of the regression model, performs residual analysis to evaluate the model's fit, and aims to provide insights into AQI changes and enhance understanding of Chennai's air quality [8, 11].

The article discusses the impact of outliers in sensor-generated data used for air pollution reduction modelling and decision-making, emphasizing the need for an incremental technique to detect outliers in temporal air quality data streams. The paper presents a methodology for assessing the effectiveness of statistical outlier detection approaches, comparing five techniques both on the entire dataset and incrementally, providing insights into the efficacy of these techniques for air quality data streams and aiding in the development of efficient air pollution control strategies [2, 13]. The paper introduces a two-stage outlier detection method for non-parametric profile monitoring to enhance the robustness of existing techniques. It uses extended least trimmed squares and a non-parametric test statistic to create an outlier identification measure, followed by hypothesis testing to identify outlying profiles, with a one-step refinement process for precise identification, demonstrating control over type-I error rates and high outlier detection power in simulations. The method is tested with real data, addressing the issue of outlying profiles in statistical process control [22, 25]. This article discusses the use of machine learning for anomaly detection to identify outliers or abnormal data points in air quality measurements. The study aims to analyze data and pinpoint pollution concentration outliers based on probability factors, contributing to the understanding of local air quality requirements [1, 18].

3 MATERIALS AND METHODS

3.1 Hybrid Model

This paper presents a hybrid strategy for outlier discovery in data, integrating K-means clustering for categorization and a mixture of Random Forest and Gradient Boosting Classifier for outlier classification (Figure: 1). The method takes advantage of the capabilities of both unsupervised and supervised learning techniques, improving the resilience and accuracy of outlier detection.

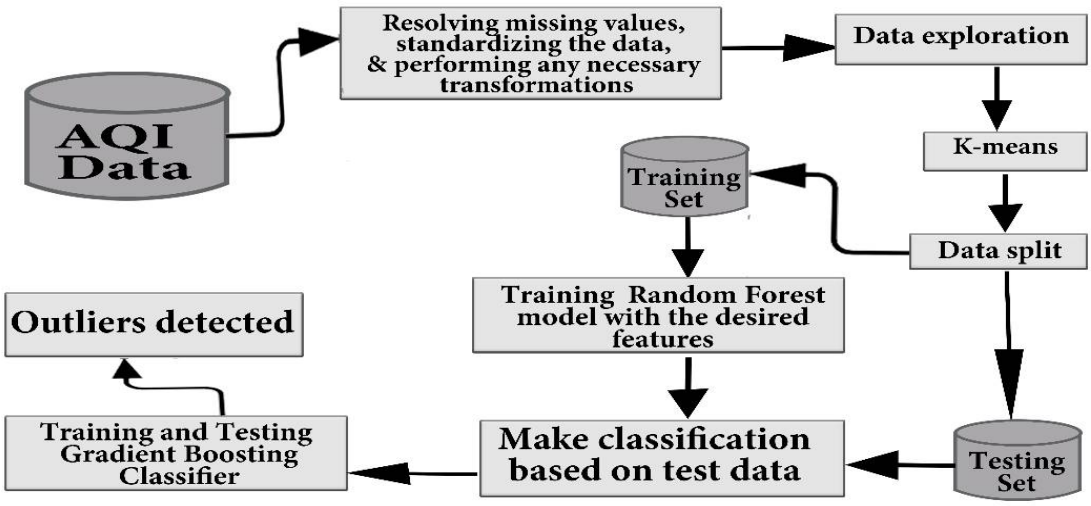

Figure 1: Flowchart of the proposed models.

3.2 Model Specification of the Hybrid

3.2.1 K-means

Given a collection of observations (vectors), divide them into K clusters so that the sum of the squared Euclidean distances between each observation and the mean of its assigned cluster (the centroid) is minimized. The goal function of K-means, also known as the inertia or within-cluster sum of squares (WCSS), is described mathematically as follows:

$$\sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{1}$$

Where K is the number of clusters, C_i represents the set of observations in the i-th cluster, x is a single observation (a vector), and μ is the centroid of the i-th cluster, defined as the mean of the vectors in C_i. Until convergence, the algorithm iteratively conducts two steps:

Step 1: Each observation is allocated to the cluster with the closest centroid, where the closest is determined by Euclidean distance.

Step 2: Recompute the centroids by taking the mean of all observations in the cluster.

3.2.2 RF - GBC

The following is the mathematical formula of RF - GBC

$$G_M(X) = G_0(X) + \sum(\vee \curlyvee_m) \tag{2}$$

Where: X is the output of Random Forest, $G_M(X)$ is the final prediction of the hybrid after M stages of boosting, $\vee$ is a learning rate ranging from 0 to 1, which governs the degree to which the new tree prediction $\curlyvee_m$ contributes to the combined prediction.

$$X = Y(x) \tag{3}$$

The final prediction Y(x) of the Random Forest is the most common class prediction among all the trees. This is also known as majority voting:

$$Y(x) = mode(y_i(x)) \tag{4}$$

Where mode is the most common prediction among all the decision trees in the forest and $y_i(x)$ represents the prediction made by the i-th decision tree in the forest. The GBC begins by predicting the average value of the target variable in the training set using a loss function, determining the loss for each instance. It then creates a new weak model that forecasts the negative gradient of the loss function concerning the target variable, approximating the error of recent forecasts. This process is repeated until a set number of weak learners have been added or the loss reaches a certain threshold, using a gradient descent approach to calculate weights for each weak learner's contribution, which are then added to the ensemble's current predictions.

1. Is the Initial model with a constant value:

$$G_0(X) = argmin\_\Upsilon \sum L(y_i, \Upsilon) \tag{5}$$

Where L is the loss function, y_i are the true target values and it is either 0 or 1 and Υ is a constant. The argmin operation finds the value Υ that minimizes the sum of the loss function over all data points. Making an initial constant forecast value of G_0 is the first step. Log loss, often known as cross-entropy loss or L, is the loss function that we are utilizing.

$$L = -(y_i \cdot log(p) + (1 - y_i) \cdot log(1 - p)) \tag{6}$$

The anticipated probability of class 1 is p. Depending on the target class y_i, you can observe L taking on different values.

$$L = \begin{cases} -log(p) & \text{if} \quad y_i = 1 \\ -log(1-p) & \text{if} \quad y_i = 0 \end{cases} \tag{7}$$

Since $-log(x)$ is a decreasing function of x, our loss will be reduced the better our prediction (by raising p for yi=1) is. $argmin$ denotes that we are looking for the gamma value Υ that minimizes $L(yi, \Upsilon)$. Even though it would be simpler to assume that Υ is the anticipated probability p, we do so because doing so simplifies all the calculations that follow. For those who missed it, the definition of log-odds, which we reviewed in the previous section, is log(odds) = log(p/(1-p)). We are converting the loss function into the log-odds function to be able to solve the arming issue in terms of log-odds.

$$L = -(y_i \cdot log(p) + (1 - y_i) \cdot log(1 - p)) \tag{8}$$

$$= -(y_i \cdot (log(p) - log(1 - p)) + log(1 - p)) \tag{9}$$

$$= -\left(y_i \cdot log\left(\frac{p}{1-p}\right) + log(1 - p)\right) \tag{10}$$

$$= -(y_i \cdot log(odds) + log(1 - p)) \tag{11}$$

Now, we may substitute anything with a log-odds expression for p in the equation above. P may be expressed by log-odds by changing the already given log-odds expression:

$$log\left(\frac{p}{1-p}\right) = log(odds) \tag{12}$$

$$\frac{p}{1-p} = e^{log(odds)} \tag{13}$$

$$p = (1 - p)e^{log(odds)} \tag{14}$$

$$\left(1 + e^{log(odds)}\right)p = e^{log(odds)} \tag{15}$$

$$p = \frac{e^{log(odds)}}{1 + e^{log(odds)}} \tag{16}$$

The preceding L equation is then simplified by substituting this value for p.

$$L = -\left(y_i \cdot log(odds) + log\left(1 - \frac{e^{log(odds)}}{1 + e^{log(odds)}}\right)\right) \tag{17}$$

$$= -\left(y_i \cdot log(odds) + log\left(\frac{1}{1+e^{log(odds)}}\right)\right) \tag{18}$$

$$= -\left(y_i \cdot log(odds) + log(1) - log\left(1+e^{log(odds)}\right)\right) \tag{19}$$

$$= -\left(y_i \cdot log(odds) - log\left(1+e^{log(odds)}\right)\right) \tag{20}$$

We are now looking for Υ that minimizes $\sum L$ (please note that we are considering it to be log odds). A derivative of $\sum L$ about log odds is being taken.

$$\frac{\partial}{\partial log(odds)} \sum_{i=1}^{n} L = -\frac{\partial}{\partial log(odds)} \sum_{i=1}^{n} \left(y_i \cdot log(odds) - log\left(1+e^{log(odds)}\right)\right) \tag{21}$$

$$= -\sum_{i=1}^{n} y_i + n\frac{e^{log(odds)}}{1+e^{log(odds)}} \tag{22}$$

$$= -\sum_{i=1}^{n} y_i + np \tag{23}$$

To make the equations above simpler, we substituted p for the fraction containing the log-odds. The next step is to put $\partial \sum L / \partial log(odds)$ equal to 0 and solve for p.

$$-\sum_{i=1}^{n} y_i + np = 0 \tag{24}$$

$$np = \sum_{i=1}^{n} y_i \tag{25}$$

$$p = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} \tag{26}$$

y is either 0 or 1 in this binary classification issue. As a result, the mean of y equals the fraction of class 1. You can probably understand why we chose $p = mean(y)$ for our first forecast.
We convert it to log-odds since Υ is log-odds rather than probability p.

$$G_0(X) = \dot{\Upsilon} = log\left(\frac{\overline{y}}{1-\overline{y}}\right) \tag{27}$$

2. For each stage m = 1 to M (where M is the total number of stages):
(a). Compute the negative gradient (also known as the residual or pseudo-residual):

$$r_{im} = -\left[\frac{\partial L(y_i, G(X_i))}{\partial G(X_i)}\right], \text{ for } i = 1 \text{ to n} \tag{28}$$

Where r_{im} is the residual for the i-th observation at the m-th stage of boosting. The equation is applied for i = 1 to n, meaning it is calculated for each data point in the datasets. In simpler terms, the equation is calculating the direction and magnitude to adjust the prediction for each observation to reduce the error (loss), hence improving the model's prediction.
Here, let's calculate the residuals.

$$r_{im} = -\frac{\partial}{\partial log(odds)} L \tag{29}$$

$$= -\frac{\partial}{\partial log(odds)} \left(y_i \cdot log(odds) - log\left(1+e^{log(odds)}\right)\right) \tag{30}$$

$$= y_i - \frac{e^{log(odds)}}{1+e^{log(odds)}} \tag{31}$$

$$= y_i - p \tag{32}$$

You may now see why we refer to r residuals. This also offers us the intriguing insight that the negative gradient that provides us with the direction and size of the loss is merely residual.

(b).Train regression tree with features X against r and create terminal node reasons R_{jm} for $j = 1,...,Jm$, j symbolizes a terminal node (or leaf) in the tree, m the tree index, and capital J the total number of leaves.

(c). Compute the multiplier Υ_m that minimizes the loss:

$$\Upsilon_{jm} = argmin\_\Upsilon \sum L(y_i, G_{m-1}(X_i) + \Upsilon), \text{ for } j = 1,...,Jm \tag{33}$$

On each terminal node j, we are trying to find the value of Υjm that minimizes the loss function. $\sum X_i \in R_{jm}L$ denotes that we are adding up all of the losses on the X_is that are connected to the terminal node R_{jm}. Now let's add the loss function to the formula.

$$\Upsilon_{jm} = argmin\_\Upsilon \sum -\left(y(G_{m-1}(X_i) + \Upsilon) - log\left(1 + e^{G_{m-1}(X_i)+\Upsilon}\right)\right) \tag{34}$$

It will be quite difficult to solve this equation for Υ_{jm}. We are approximating L with a second-order Taylor polynomial to make it more easily solvable. A Taylor polynomial is a method for representing any function as a polynomial with an infinite/finite number of terms. While we will not go into depth here, if you are interested, you may look at this tutorial which explains the concept beautifully. The second-order Taylor polynomial is used to approximate L:

$$L(y, G_{m-1}(X_i) + \Upsilon) \approx L(y, F_{m-1}(X_i)) + \frac{\partial}{\partial G} L(y_i, G_{m-1}(X_i))\,\Upsilon + \frac{1}{2}\frac{\partial^2 L(y_i, G_{m-1}(X_i))}{\partial G^2}\Upsilon^2 \tag{35}$$

We are seeking the value of jm that causes the derivative of $\sum$(*) to equal zero by substituting this approximation for L in the equation of Υ_{jm}.

$$\frac{\partial}{\partial \Upsilon} \sum \left(L(y_i, G_{m-1}(X_i)) + \frac{\partial}{\partial G} L(y_i, G_{m-1}(X_i))\,\Upsilon + \frac{1}{2}\frac{\partial^2 L(y_i, G_{m-1}(X_i))}{\partial G^2}\Upsilon^2\right) = 0 \tag{36}$$

$$\sum \left(\frac{\partial}{\partial G} L(y_i, G_{m-1}(X_i)) + \frac{\partial^2 L(y_i, G_{m-1}(X_i))}{\partial G^2}\Upsilon\right) = 0 \tag{37}$$

$$\sum \frac{\partial^2 L(y_i, G_{m-1}(X_i))}{\partial G^2}\Upsilon = -\sum \frac{\partial}{\partial G} L(y_i, G_{m-1}(X_i)) \tag{38}$$

$$\Upsilon = \frac{-\sum \frac{\partial}{\partial G} L(y_i, G_{m-1}(X_i))}{\sum \frac{\partial^2}{\partial G^2} L(y_i, G_{m-1}(X_i))} \tag{39}$$

As $\partial L/\partial G$ was previously calculated in the step below:

$$\frac{\partial L(y_i, G(X_i))}{\partial G(X_i)} = -(y_i - p) \tag{40}$$

This is being used as a substitution for $\partial L/\partial G$ in the Υ equation.

$$\Upsilon = \frac{\sum (y_i - p)}{\sum \frac{\partial}{\partial G}(-y_i + p)} \tag{41}$$

$$= \frac{\sum (y_i - p)}{\sum \frac{\partial}{\partial G}\left(-y_i + \frac{e^{log(odds)}}{1+e^{log(odds)}}\right)} \tag{42}$$

$$= \frac{\sum (y_i - p)}{\sum \frac{\partial}{\partial G}\left(-y_i + e^{log(odds)}\left(1 + e^{log(odds)}\right)^{-1}\right)} \tag{43}$$

$$= \frac{\sum(y_i - p)}{\sum \frac{\partial}{\partial G}\left(e^{log(odds)}\left(1+e^{log(odds)}\right)^{-1}\right)} \tag{44}$$

$$= \frac{\sum(y_i - p)}{\sum \left(e^{log(odds)}\left(1+e^{log(odds)}\right)^{-1} - e^{2log(odds)}\left(1+e^{log(odds)}\right)^{-2}\right)} \tag{45}$$

$$= \frac{\sum(y_i - p)}{\sum \left(\frac{e^{log(odds)}}{1+e^{log(odds)}} - \left(\frac{e^{log(odds)}}{1+e^{log(odds)}}\right)^2\right)} \tag{46}$$

$$= \frac{\sum(y_i - p)}{\sum \left(p - p^2\right)} \tag{47}$$

$$= \frac{\sum(y_i - p)}{\sum p(1-p)} \tag{48}$$

This simplified equation, which was utilized in the preceding section, is what finally arrived at the value of Υ_{jm}.

(d). Update the model:

$$G_m(X) = G_{m-1}(X) + \vee \sum \Upsilon_{jm} 1(X \in R_{jm}) \tag{49}$$

The prediction of the combined model G_m is updated in the last phase. $\Upsilon_{jm} 1(X \in R_{jm})$ denotes that choose the value Υ_{jm} if a given X falls in a terminal node R_{jm}. Because all of the terminal nodes are exclusive, each given single x may only fall into a single terminal node, and the matching Υ_{jm} is added to the prior prediction G_{m-1} before making the updated prediction G_m. As noted in the preceding section, $\vee$ is a learning rate ranging from 0 to 1, which governs the degree to which the new tree prediction Υ contributes to the combined prediction G_m. A lower learning rate minimizes the influence of further tree prediction, but it also reduces the likelihood of the model over-fitting to the training data.

4 Data Description and Experiment

4.1 Data collection and preprocessing

This study utilizes a comprehensive dataset detailing daily air quality in Shanghai, including measures of pollutants like PM2.5, PM10, O3, NO2, CO, and the AQI. The dataset spans several years, providing a detailed historical record of air quality (Table 1). The research aims to detect outliers in the AQI and gain insights into air quality patterns, correlations between pollutants, and factors impacting the AQI, which can inform policy decisions, guide future research, and raise awareness about air quality issues. The initial data in this experiment are handled as follows:

Table 1: Initial Air Quality Data

Date	Pollutant Concentrations (μg/m3)					AQI
	PM2.5	PM10	O3	NO2	CO	
2021-01-31	58.0	56.0	37.0	17.0	7.0	162.0
2021-01-30						
2021-01-29	146.0	62.0	24.0	13.0	6.0	320.0
2021-01-28	132.0	72.0	28.0	20.0	10.0	309.0
2021-01-27	90.0	50.0	38.0	16.0	9.0	200.0

4.1.1 Data filling

Data entry. Data loss during the air quality data collection process may be brought on by network outages, storage problems, and other issues, like the data on January 30, 2021. The learning effect of the model will be impacted by this poor-quality data. Due to the issue of missing values in the original data, the final forecast accuracy is not very good. This experiment utilized the average value of the data from one hour before and one hour after to fill in the missing parts (Table 2), as indicated in (formula 50), taking into account the fact that the air pollution data typically varies steadily with time and there is typically no rapid shift in values [20].

$$V_t = \frac{V_{t-1} + V_{t+1}}{2} \tag{50}$$

Where V_t represents the missing value at time t, V_{t-1} represents the data from one day before to time t, and V_{t+1} represents the data from one day after time t. The six pollution indices included in this experiment's data set are PM2.5, O3, NO2, PM10, and CO, which are used by the environmental protection agency to compute AQI [24, 26].

Table 2: Inputting Air Quality Data

Date	Pollutant Concentrations (μg/m3)					AQI
	PM2.5	PM10	O3	NO2	CO	
2021-01-31	58.0	56.0	37.0	17.0	7.0	162.0
2021-01-30	102	59	30.5	15	6.5	241
2021-01-29	146.0	62.0	24.0	13.0	6.0	320.0
2021-01-28	132.0	72.0	28.0	20.0	10.0	309.0
2021-01-27	90.0	50.0	38.0	16.0	9.0	200.0

4.1.2 Data normalization

The sample values of some features in the data set deviate significantly from those of other features, which might cause sluggish convergence and lower model training accuracy. In this experiment, the original data are processed using z-score normalization as given in (formula 51), where σ is the original data standard deviation, $\overline{X}$ is its mean, and X^* is the value after standardization. The data is dimensionless and scaled to the same interval after data normalization. Additionally, because the features are comparable and the trend and relative size of the scaled data remain constant, the model convergence occurs more quickly.

$$X^* = \frac{X - \overline{X}}{\sigma} \tag{51}$$

The indicators listed below are used to measure the AQI (Table 3).

Description	**AQI Value**	**PM10**	**PM2.5**	**CO**	**O3**	**NO2**
Good + Satisfactory	0-100	0-100	0-60	0-1.7	0-50	0-43
Moderate	101-200	101-250	61-90	1.8-8.7	51-84	44-96
Poor	201-300	251-350	91-120	8.-14.8	85-104	97-149
Very Poor	301-400	351-430	121-250	14.9-29.7	105-374	150-213
Severe	401-500	431-550	251-350	29.8-40	375-450	214-750

Table 3: AQI Levels and Associated Pollutants

4.2 Exploratory Data Analysis

In this study, we conducted an exploratory data analysis to gain a deeper understanding of the AQI dataset. We visualized the data in various ways to understand the distribution of AQI values, the trend of AQI over time, and the relationships between AQI and other pollutants. The daily AQI showed significant fluctuations, with periods of both high and low AQI, suggesting possible seasonal patterns. However, discerning a clear trend from the daily AQI was challenging due to its high variability. A clearer picture of the overall trend was provided by the average annual AQI (Figure 2). The histogram of AQI values revealed that most values clustered in the 100 to 300 range, with fewer values in the upper range, indicating a right-skewed distribution. This suggests that certain days experience extremely poor air quality. The box plot confirmed the observations from the histogram, providing a summary of the statistical distribution of AQI. The box represented the interquartile range (IQR), the middle 50% of the data, with the median represented by the line in the centre of the box. The "whiskers" extended to the minimum and maximum values within 1.5 times the IQR, and outliers, which fell outside of this range, were plotted as separate points. A few outliers were observed in the higher range (Figure 3).

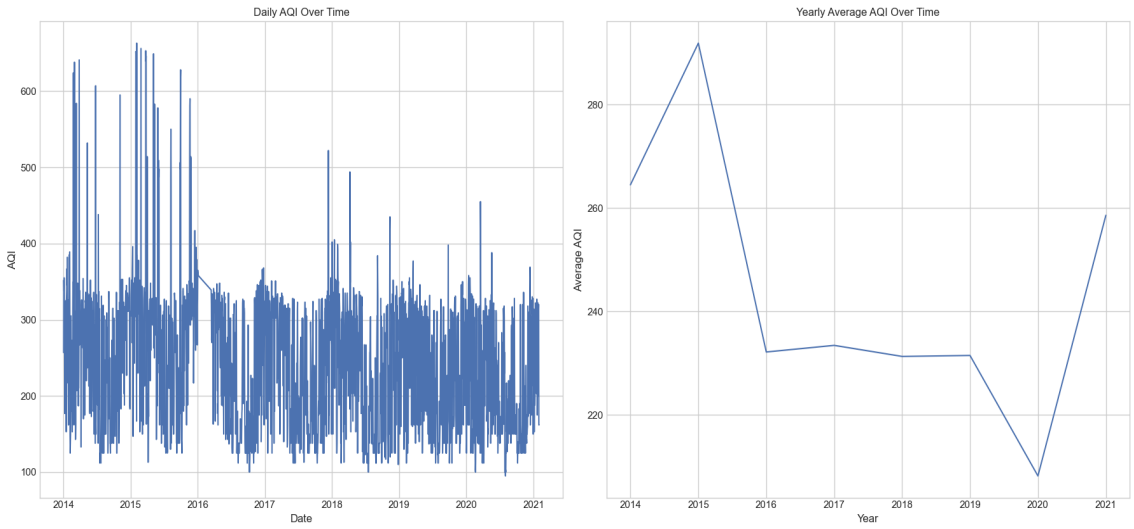

Figure 2: AQI Trend by Day and Year.

AQI versus PM2.5: This graph illustrates a favourable relationship between AQI and PM2.5 levels. This suggests that greater levels of PM2.5 (particulate matter with a diameter of fewer than 2.5 micrometres) are related to a higher AQI. This positive link is predicted given that PM2.5 is a key contaminant that affects air quality. AQI versus PM10: This plot, too, indicates a favourable relationship between AQI and PM10 levels (particulate matter smaller than 10 micrometres in diameter). This implies that greater PM10 levels are related to a higher AQI. PM10, like PM2.5, is a substantial pollutant, and hence a positive connection with AQI is predicted (Figure 4).

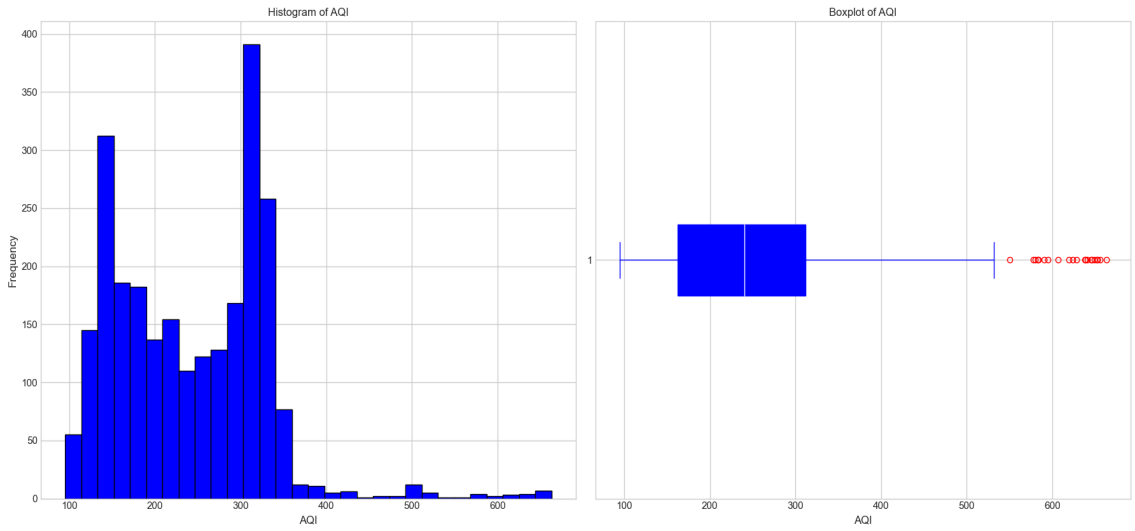

Figure 3: Histogram and Box plot of AQI.

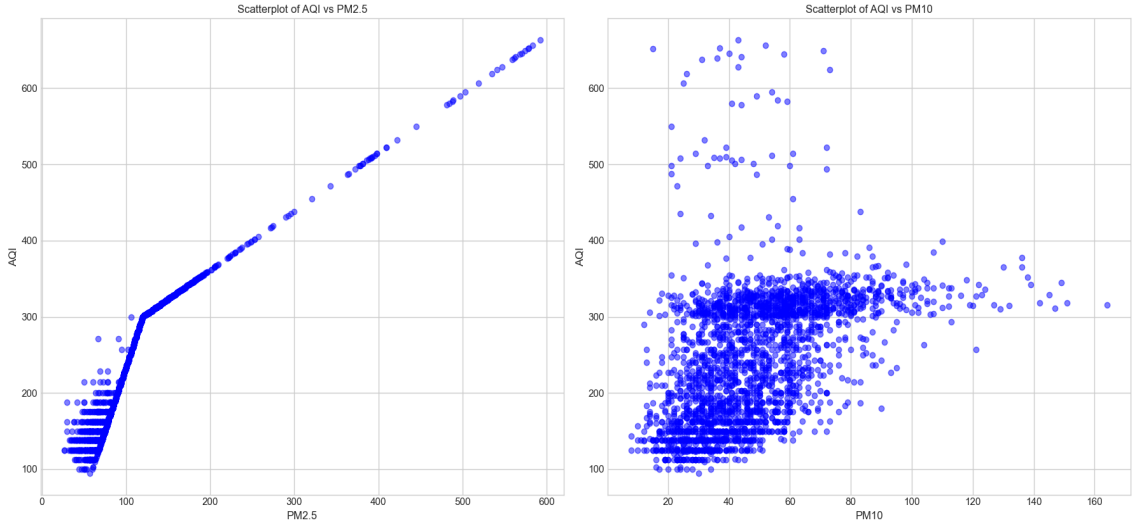

Figure 4: Scatter plots of AQI vs Other Pollutants.

4.3 Evaluation Metrics

To assess the efficacy of a classification model, several metrics such as accuracy (ACC), precision, recall, Receiver Operating Characteristic (ROC) curve, and F1-score can be used. Before delving into each of these measurements, it's critical to grasp the following abbreviations: TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). These formulae most effectively reflect the concepts that are examined as a consequence of a classifier.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{52}$$

$$Precision = \frac{TP}{TP + FP} \tag{53}$$

$$Recall = \frac{TP}{TP + FN} \tag{54}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{55}$$

5 RESULTS AND DISCUSSION

In this study, machine learning algorithms were used to generate labels for the AQI and categorize them accordingly. The data was divided into two clusters based on commonalities in pollution measurements, with the majority of data points assigned to Cluster 0 having 2453 and a smaller number to Cluster 1 having 50. Both the RF and GBC models performed flawlessly across several metrics, including accuracy, precision, recall, and the F1 score (Table 4). The Receiver Operating Characteristic (ROC) curves for the models were visualized, with the Area Under the Curve (AUC) score calculated for each model (Figure 5). The RF-GBC models appeared to outperform the other variants, indicating successful discrimination between the two groups and balanced trade-offs between sensitivity and specificity.

Table 4: Performance evaluation summary

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.96	0.93	0.91	0.92
SVM	0.95	0.92	0.89	0.91
KNN	0.94	0.88	0.88	0.88
Naive Bayes	0.96	0.92	0.94	0.93
RF-GBC	0.99	0.99	0.96	0.98

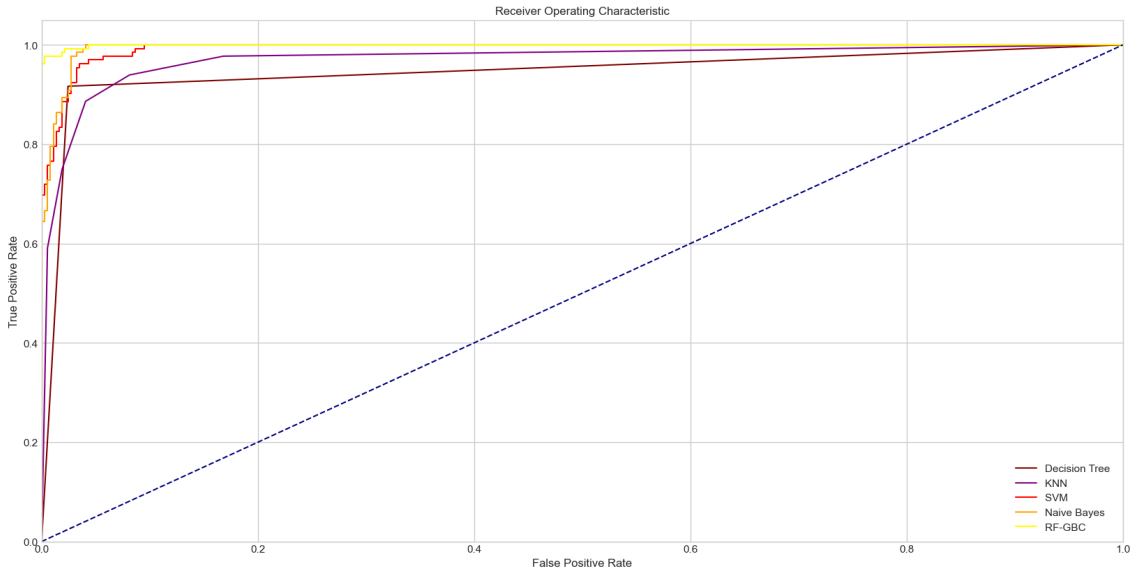

Figure 5: AUC-ROC cuve for Ensemble RF-GBC.

5.1 Discussion

In this study, we explored the relationship between the AQI and various pollutants using a comprehensive dataset from Shanghai. Through exploratory data analysis, we observed the distribution of AQI values, analyzed AQI trends over time, and investigated correlations between AQI and other pollutants. Our analysis revealed a slight downward trend in AQI over time, suggesting a potential improvement in air quality. However, the existence of numerous outliers at the higher end of the AQI spectrum indicates occasional instances of extremely poor air quality. Notably, we found positive correlations between AQI and PM2.5 and PM10, indicating that higher levels of these pollutants correspond to a higher AQI. In the machine learning phase of our study, we utilized K-Means clustering labels as the basis for several classification models, including RF, GBC, Decision Tree, SVM, KNN, and Naive Bayes. Evaluation metrics such as accuracy, precision, recall, and F1 score revealed near-perfect or perfect scores for most models, suggesting their effectiveness in classifying AQI into two distinct groups. However, our study is not without limitations. The K-means labels are somewhat arbitrary and may not accurately represent the underlying categories in the AQI. The strong performance of the models may be attributed more to the clear separability of the two clusters rather than their predictive power. Additionally, the imbalance in our dataset could potentially affect model performance. Despite these limitations, our study provides valuable insights into Shanghai's air quality and the factors influencing AQI. It also demonstrates the potential of machine learning in analyzing and predicting air quality. Future research could delve deeper into the factors contributing to poor air quality, validate our findings using other datasets, and refine the models to enhance their robustness and reliability.

5.2 Conclusion

In this research, we have effectively employed machine learning techniques to delve into the complexities of the Air Quality Index (AQI) in Shanghai. Our focus has been on the detection of outliers, which are crucial in understanding the full spectrum of air quality conditions and the impact of various pollutants on AQI. Through exploratory data analysis, we have unearthed significant patterns and correlations between AQI and other pollutants. This foundational understanding of the dataset has been instrumental in our subsequent analyses. We have observed a slight but promising decrease in AQI over time, suggesting a potential trend towards improved air quality. However, the existence of outliers serves as a stark reminder of the occasional instances of extremely poor air quality. We have utilized the power of K-Means clustering to create distinct AQI categories, which have then been used to train a suite of classification models. These models, including but not limited to RF - GBC but Decision Tree, Support Vector Machine, KNN, and Naive Bayes, have demonstrated exceptional performance, achieving near-perfect or perfect scores across multiple performance metrics. Our research stands as a testament to the transformative power of machine learning in shaping our understanding of air quality. By leveraging these insights and predictions, we aspire to inform and influence air quality management and policy decisions, ultimately contributing to the creation of healthier, safer, and more sustainable environments.

CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT
Data is available: https://www.kaggle.com/datasets/erhankul/shanghai-air-pollution-and-wheather-20142021

References

[1] T.R.V. Anandharajan, K.K. Vignajeth, G. Abhishek Hariharan, and R. Jijendiran. Identification of outliers in pollution concentration levels using anomaly detection. mar 2016.

[2] Fabrizio Angiulli, Fabio Fassetti, Giuseppe Manco, and Luigi Palopoli. Outlying property detection with numerical attributes. *Data Mining and Knowledge Discovery*, 31(1):134–163, mar 2016.

[3] Shin Araki, Hikari Shimadera, Kouhei Yamamoto, and Akira Kondo. Effect of spatial outliers on the regression modelling of air pollutant concentrations: A case study in japan. *Atmospheric Environment*, 153:83–93, mar 2017.

[4] A. Bhushan, M. H. Sharker, and H. A. Karimi. INCREMENTAL PRINCIPAL COMPONENT ANALYSIS BASED OUTLIER DETECTION METHODS FOR SPATIOTEMPORAL DATA STREAMS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-4/W2:67–71, jul 2015.

[5] X. Chuanqi, Z. Zhi, and L. Guangjiu. Air pollutant spatiotemporal evolution characteristics and effects on human health in north china. *Chemosphere*, 294, 2022.

[6] Kruti Davda. Air quality index – importance of aqi, 2019.

[7] Song Feng, Qi Hu, and Weihong Qian. Quality control of daily meteorological data in china, 1951–2000: a new dataset. *International Journal of Climatology*, 24(7):853–870, may 2004.

[8] Hamid Ghorbani. MAHALANOBIS DISTANCE AND ITS APPLICATION FOR DETECTING MULTIVARIATE OUTLIERS. *Facta Universitatis, Series: Mathematics and Informatics*, page 583, oct 2019.

[9] I. Hossain, S. Rahman, and S. Sattar. Environmental overview of air quality index (aqi) in bangladesh: Characteristics and challenges in present era. *Int. J. Res. Eng. Technol.*, 4, 2021.

[10] M. Ikram and Z. J. Yan. Statistical analysis of the impact of aqi on respiratory disease in beijing: Application case 2009. *Energy Proc.*, 107, 2017.

[11] A. Loganathan, P. Sumithra, and V. Deneshkumar. Estimation of air quality index using multiple linear regression. *Applied Ecology and Environmental Sciences*, 10(12):717–722, dec 2022.

[12] Manish Mahajan, Santosh Kumar, Bhasker Pant, and Rijwan Khan. Improving accuracy of air pollution prediction by two step outlier detection. feb 2021.

[13] Manish Mahajan, Santosh Kumar, Bhasker Pant, and Umesh Kumar Tiwari. Incremental outlier detection in air quality data using statistical methods. oct 2020.

[14] Mohsin Munir, Shoaib Ahmed Siddiqui, Muhammad Ali Chattha, Andreas Dengel, and Sheraz Ahmed. FuseAD: Unsupervised anomaly detection in streaming sensors data by fusing statistical and deep learning models. *Sensors*, 19(11):2451, may 2019.

[15] Mingfei Niu, Kai Gan, Shaolong Sun, and Fengying Li. Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead PM 2.5 concentration forecasting. *Journal of Environmental Management*, 196:110–118, jul 2017.

[16] Junhyeok Park, Youngsuk Seo, and Jaehyuk Cho. Unsupervised outlier detection for time-series data of indoor air quality using LSTM autoencoder with ensemble method. *Journal of Big Data*, 10(1), may 2023.

[17] Maurras Ulbricht Togbe, Yousra Chabchoub, Aliou Boly, Mariam Barry, Raja Chiky, and Maroua Bahri. Anomalies detection using isolation in concept-drifting data streams. *Computers*, 10(1):13, jan 2021.

[18] Daniel W. Urish. The practical application of surface electrical resistivity to detection of groundwater pollution. *Ground Water*, 21(2):144–152, mar 1983.

[19] V. M. van Zoest, A. Stein, and G. Hoek. Outlier detection in urban air quality sensor networks.

[20] J. Wang, J. Li, X. Wang, J. Wang, and M. Huang. Air quality prediction using ct-lstm. *Neural Comput. Appl.*, 33, 2020.

[21] Jianzhou Wang, Pei Du, Yan Hao, Xin Ma, Tong Niu, and Wendong Yang. An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting. *Journal of Environmental Management*, 255:109855, feb 2020.

[22] Tao Wang, Yunlong Wang, and Qingpei Zang. Outlier detection in non-parametric profile monitoring. *Statistics*, 56(4):805–822, jun 2022.

[23] Huangjian Wu, Xiao Tang, Zifa Wang, Lin Wu, Miaomiao Lu, Lianfang Wei, and Jiang Zhu. Probabilistic automatic outlier detection for surface air quality measurements from the china national environmental monitoring network. *Advances in Atmospheric Sciences*, 35(12):1522–1532, oct 2018.

[24] Y. Xie, B. Zhao, L. Zhang, and L. Rong. Spatiotemporal variations of pm(2.5) and pm(10) concentrations between31 chinese cities and their relationships with so2, no2, co and o3. *Particuology.*, 20, 2015.

[25] Guan Yu, Changliang Zou, and Zhaojun Wang. Outlier detection in functional observations with applications to profile monitoring. *Technometrics*, 54(3):308–318, aug 2012.

[26] Z. Yu, D. S. Moirangthem, and L. Minho. Continuous tirmsecale long-short term memory neural network for human intent understanding. *Front. Neurorobot.*, 11, 2017.

[27] Xiaoxia Zhang and Hao Gan. An enhanced anomalies detection method based on isolation forest and fuzzy set. nov 2022.

[28] Yang Zhang, Nirvana Meratnia, and Paul Havinga. Outlier detection techniques for wireless sensor networks: A survey.

[29] Arthur Zimek and Peter Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *WIREs Data Mining and Knowledge Discovery*, 8(6), aug 2018.