Ensemble-SMOTE Model to Evaluate Air Quality in the Industrial Area in Chavara

Susymary Johnson[1], Deepalakshmi Perumalsamy[2]

[1]*Research Scholar, Department of Computer Applications, Kalasalingam Academy of Research and Education, India, susymaryj@gmail.com*
[2]*Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, India, deepa.kumar@klu.ac.in*

Abstract: Air quality is a critical environmental concern, particularly in industrial areas where emissions from factories can significantly impact the health of nearby populations. This study focuses on evaluating the air quality on pollutants like SO2, NO2, PM10, and SPM in the Kerala Minerals and Metals Limited (KMML) industrial area in Chavara, Kerala, India. To predict air quality indicators accurately, the researchers used a combination of artificial intelligence techniques. By comparing error metrics across different approaches, they identified the optimal method for accurate predictions. The study employed machine learning algorithms and SMOTE to predict Air Quality Index (AQI) levels. The ensemble SMOTE method outperformed individual classifiers like KNN, SVM, DT, RF, and GaussianNB, achieving higher accuracy, precision, recall, and F1-score, indicating its effectiveness in predicting AQI levels. The study also highlighted the importance of data preprocessing and balancing for improved prediction accuracy.

Keywords: Air Quality Evaluation, Machine Learning Techniques, Industry 4.0, and Chavara Kerala.

1. Introduction

The emissions from combustion processes, such as those in factories and cars, are the most common cause of outdoor air pollution from human-caused causes. The industrial area's monitoring of air quality system, as well as the neighboring residential and commercial districts, may be useful. They aid in the validation of the consequences of industrial pollution in metropolitan regions with such dense populations. Such a system is beneficial not only for the industry's environmental impact assessment but also for regulatory air quality monitoring, gas leaks, and fugitive emissions. Industrial emission tracking may be done effectively and economically using Oizom's Fenceline Monitoring technology. The air pollution in the neighborhood was worsened since these industrial sectors released a lot of harmful substances into the air. Young people's respiratory health may deteriorate due to the area's closeness to the industry and the length of time spent there. The effects of air pollution on children were found to be more severe than those on adults [1], [2]. The influence of city-region spatial structures on pollution levels is a growing concern in the field of sustainable regional development as the city-region development plan gains traction. Few studies have examined the impact of industrial pollution reduction on environmental quality from either a polycentric or monocentric vantage point, despite the fact that prior research has focused on the spatial structure influencing

environmental quality [3].

Sustainable economic growth is significantly impacted by industrial agglomeration, which alters a city's economic structure, strategic architecture, and resource status. A critical challenge for the nation's high-quality economic growth is the connection among industrial agglomeration, air pollution, and long-term sustainability. As a growing nation, China has a history of serious air pollution [4].

Among the most pressing environmental issues of our day is air pollution. Bulgaria is no different from any other country in that industry is the primary source of anthropogenic emissions. Nitrogen dioxide (NO2), carbon monoxide (CO), methane (CH4), and sulfur dioxide (SO2) are the four most major air pollutants that affect air quality and contribute to climate change, either directly or indirectly [5].

Air pollution has a significant impact on people's health and everyday lives. To reduce air pollution, the Chinese administration is stepping up its efforts. An instrument for resolving environmental and economic growth-related problems is the improvement of industrial structure [6]. The sustainable growth of China has been severely impeded by air pollution. Given the fast-paced digital economy, it is still not obvious how industrial upgrading affects air pollution. The effects of digital economy-driven industrial structure upgrades on air pollution must be studied [7]. Air quality monitoring in urban, rural, or industrial settings is one of many relevant uses for the rapidly developing field of unmanned aerial vehicles. Their main concern is tracking the spread of air pollution once its origin is identified. Because several chimneys in an industrial region release smoke, the source of the pollution is frequently unclear and hard to determine [8], [9], [10]. Some of the air pollutants are follows,

Particulate Matter (PM10 and PM2.5): These particles can penetrate deep into the lungs and bloodstream, causing respiratory and cardiovascular issues.

Sulfur Dioxide (SO2): SO2 can irritate the respiratory system and aggravate conditions like asthma.

Nitrogen Oxides (NOx): NOx can contribute to the formation of ozone and fine particulate matter, leading to respiratory problems.

The proposed method offers a novel approach by combining SMOTE for class imbalance correction with ensemble learning techniques, enhancing the accuracy and robustness of Air Quality Index (AQI) prediction models.

Contribution of the work:

The work presents a novel approach that combines SMOTE for class imbalance correction with ensemble learning techniques for predicting Air Quality Index (AQI) levels. This approach improves prediction accuracy, robustness, and generalization performance compared to existing methods, offering a more effective solution for air quality monitoring and management.

2. Related Work

For the purpose of predicting and forecasting air pollution and quality in specific areas, provide an effective combination method in [11] that makes use of the advantages of statistical and machine learning approaches. The results of that study also show that the accuracy of predictions differs across various cities and regions in India. Based on a wide range of surface and atmospheric variables such as wind velocity, air temperature, pressure, etc., they used time series analysis, regression, and Ada-boosting to predict yearly PM 2.5 concentration levels at many sites around Hyderabad. Authors used a dataset from Kaggle, ran trials using the suggested strategy, and compared the results graphically.

A fast way to identify indoor air pollution using machine learning and laser-induced breakdown spectroscopy was suggested in a study [12]. Changes in interior air quality were often caused by four typical scenes: burning carbon, incense, perfume spray, and a hot shower.

The experiment included two steps: measuring spectra and analyzing the results using an algorithm. In addition, the suggested approach demonstrated sensitivity to air components and successfully differentiated between various aerosol types. Singular values were eliminated out in that research since the signal was separated by the forest. In the meanwhile, the air environment was identified with a 99.2% success rate using the K-Nearest Neighbor method, and the spectra of various situations were evaluated using principal component analysis. To further enhance the accuracy and resilience of the interior environment, a back propagation neural network was implemented based on the construction of a high-precision the quantitative identification model.

Finding the best approach for AQI forecasting to aid in climate management is the goal of the [13] study. The best answer may be found by improving upon the most successful strategy. To get the best answer for the air quality issue, that paper's work includes extensive study and the use of new approaches like SMOTE. Their study also aims to exhibit and illustrate the precise measurements used in a manner that is instructive and illuminating, allowing for accurate comparisons and the assistance of future researchers. For the purpose of that proposal, the air quality indexes of New Delhi, Bangalore, Kolkata, and Hyderabad were calculated using three separate methods: random forest regression, support vector regression, and catalyst regression. Upon comparing the outcomes of imbalanced datasets, it was discovered that random forest regression yields the best results in Bangalore (0.5674), Kolkata (0.1403), and Hyderabad (0.3826). It also outperforms SVR and CatBoost regression in terms of accuracy for Kolkata (90.9700%) and Hyderabad (78.3672%), while CatBoost regression yields the lowest RMSE value in New Delhi (0.2792) and the best results for both Bangalore (68.6860%) and New Delhi (79.8622%).

To improve the accuracy of their prediction model, the authors of the aforementioned research used a machine learning technique to apply predictive analytics[14]. A prediction model is built to foretell future values after trends and patterns are examined using past time series data. They propose the Air Quality Forecasting Model, which makes use of these prediction models, to implement their strategy. The outcome of that model is a prediction model that, given the available data, can reliably forecast the AQI. In order to get that data, they will be employing a web scraping approach to get it from the CPCB website. Long Short-Term Memory emerged as the clear winner in the ML/DL accuracy comparison for air quality measurement across all three parameters. The data is then analyzed based on the LSTM model's anticipated AQI.

The concentrations of air pollutants in the Delhi area were measured daily and hourly in the research [15]. Various methodologies were used for that analysis. Months, seasons, and geographical features of several stations are used to conduct a comparison study. Researchers also looked at how the COVID-19 lockout affected the decrease of pollution levels. By analyzing the existing data, they can see how certain contaminants are dependent on one another, how they relate to meteorological conditions, and how the stations are correlated with one another. The air quality forecasts were created using a variety of machine learning models, including Decision Tree Regression, Linear Regression, Random Forest, Gradient Boosting Machine, and Vector Auto Regression. These models were evaluated based on their MAE, RMSE, and MAPE scores. That research aims to provide light on the seriousness of Delhi's air pollution, its root causes, and the effectiveness of planned lockdowns in reducing emission levels. The ability of Decision Tree Regression and Linear Regression models to forecast air quality across various time periods is also shown.

Using machine learning and improved secondary data modeling, a new approach to air quality prediction is presented in [16]. The dataset used includes primary meteorological conditions and pollutant concentration forecasts as well as actual meteorological measurements and observations covering the period from 23 July 2020 to 13 July 2021. The data comes from long-term air quality projections made at different monitoring stations in Jinan, China. The first

step was to conduct a thorough correlation study in order to identify 10 meteorological components. These elements would include both observed and predicted data from five different categories. Afterwards, a random forest method was used in conjunction with univariate as well as multivariate significance analyses to rank the importance of these 10 variables according to their effect on various pollutant concentrations. The concentrations of six important air pollutants were found to be significantly affected by humidity, temperature, air pressure, and general meteorological conditions throughout the year, according to seasonal characteristic analysis.

Data on annual concentrations of air pollutants are retrieved from the Pune Smart City office, the SPV responsible for carrying out the Smart City Mission in Pune City, in the study cited as [17]. Data analytics tools like Tableau as well as Machine Learning decision tree algorithms are used for analysis and visualization in that study, which focuses on pre-processing one year's worth of data pertaining to concentration levels of main AQI contaminants. Data may be better understood with the help of Tableau graphics and characteristics from correlation matrices. In order to provide the most accurate predictions about the air quality, they used a supervised machine learning technique called Random Forest in conjunction with Tableau's Time Series forecast model. Tableau dashboard trends make the changes over time easy to comprehend. The city may become "smart" thanks to data analysis, which allows for the prediction of potential air pollution level and the implementation of preventative actions by both citizens and government.

The article [18] suggests using models based on images and deep learning to predict air pollution levels. For the purpose of assessing the air quality, the simulation compiles feature data from landscape images captured by mobile devices. In order to examine how the general population views the air quality, researchers surveyed 257 individuals. Using structural analysis, the Smartpls technique may be used to find out how each variable affects the other factors and how much of an impact they have on the overall sense of air quality. With the use of image-based information and machine learning methods, that research intends to establish a new method for predicting air quality. In order to forecast the air quality index, the study applied convolutional neural networks to extract information from photos. Datasets collected from the city's network of air quality monitors were used in the research. When compared to more conventional approaches, the study's findings demonstrated that the suggested strategy might provide more precise estimates of air quality.

In order to resolve the Air Quality Analysis in that research utilizing machine learning methods, a thorough and practical strategy is given in [19]. The Air Quality Index for India is determined using data collected from many weather stations and presented in that article. Back propagation neural networks, Support vector machines, and decision trees are some of the machine learning algorithms that they are using on the collected data.

Air quality data may be accurately and quickly analyzed with the use of an IoT-based system that employs machine learning, as shown in the research [20]. In order to understand the present environmental situation, the system gathers data from a system of sensors that measure different air quality metrics. The data is then processed through ML algorithms to find patterns and make predictions about the future. The results demonstrated that the study region's air quality was negatively correlated with emissions, and the system attained an accuracy level of 0.978. With any luck, that research will lead to real-time air quality regulation and more precise interpretation of air quality data.

3. Proposed Work

Study Area

The subject area comprises 29 wards from two panchayats in the Chavara industrial area of the Kollam district, historically a hub of many large and medium-scale industries. KMML is a significant industrial establishment in Chavara, involved in the extraction of minerals and the

production of titanium dioxide. KMML is the only integrated Titanium Dioxide facility in India that includes mining, mineral separation, synthetic rutile, and pigment production plants. It produces very pure rutile Titanium Dioxide through chloride treatment, resulting in Titanium Tetra Chloride as a byproduct, which is highly reactive with water. The waste dump primarily consists of chlorine, which, although highly reactive and not persistent in the environment, can increase levels of NO2, SO2, PM10, and SPM in the atmosphere. This study aims to analyze air quality using advanced machine learning techniques for more accurate and timely insights. Fig 1 is the depiction of the study area.

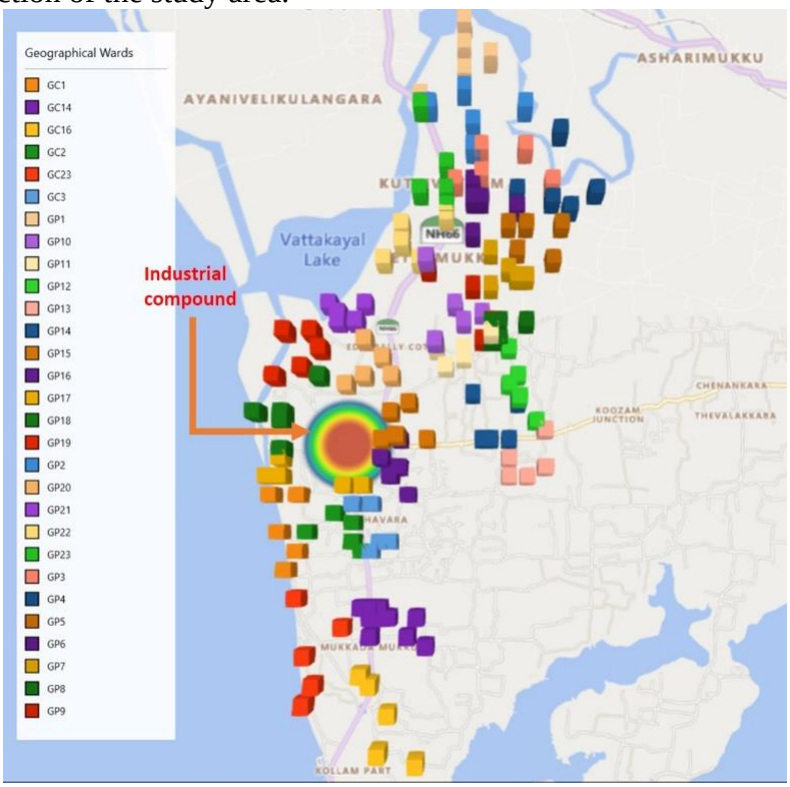

Fig 1. Depiction of the Study Area

Data Collection

In this study, we focus on data collected from the only air pollution monitoring station situated within the industrial compound. Daily monitored data from this station has been obtained from the Kerala Pollution Control Board, spanning from January 2010 to September 2022. The pollutants monitored include SO2, NO2, PM10, and total SPM. The concentration distribution graphs indicate that levels of these pollutants were particularly high and varied significantly between 2011 and 2016, potentially due to chlorine leaks in 2011 and 2014. Fig. 2 shows the concentration distribution graphs of the pollutants.

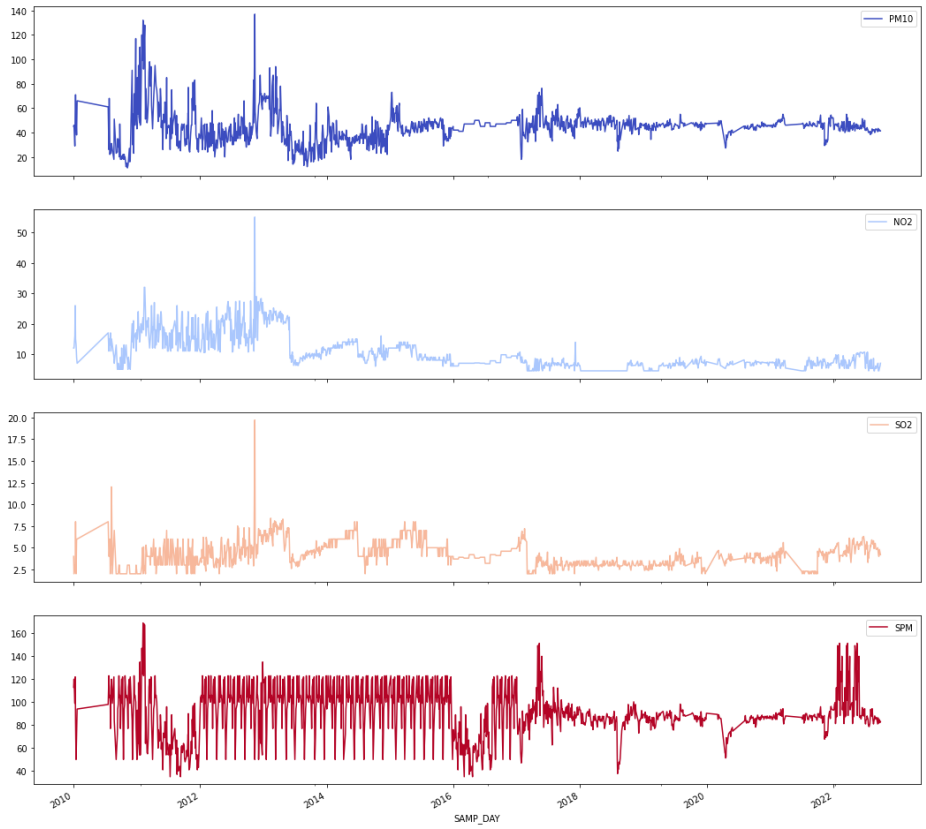

Fig. 2: The concentration distribution graphs of the Pollutants

As a second source, we have collected the air pollution data from the literature, stating one of the air pollution measurement stations as within 0.5 kilometers from the East of the industrial compound and another 14 kilometers away from the North of the industrial compound [21]. The specified data has recorded in the year 2011. We were able to find considerable difference in the KMML station data and the average of the literature data in and around the KMML station in 2011. PM10 is 4.479 times, SPM is 5.37 times, SO2 is 20.287 times, and NO2 is 2.352 times higher than the actual KMML station data. Fig 3 depicts the data difference. Using this literature data, we replaced the data for each pollutant in the Chavara monitoring station for the years 2010 to 2022 by a linear trend determined by the closest known values [22], [23].

STATION	LONG	LAT	DIST_FROM_IND	RSPM(PM10)_CONC	TSPM(SPM)_CONC	SO2_CONC	NO2_CONC
S1	76.534	8.996	0.5	278.75	422.417983	107.228	48.9
S2	76.536	9.01	1	231.25	340.8262454	59.552	31.8
S3	76.538	9.001	0.5	248.75	404.3438639	84.392	34.7
S4	76.541	9.02	3	263.125	425	88.232	57.5
S4	76.539	8.995	1	257.5	411.5735115	90.34	51.448
S6	76.518062	9.120766	14	235.625	343.4082625	19.052	17.298
KMML_CALCULATED	76.53198	9.00066	0	252.5	391.2616444	74.79933333	40.27433333
KMML_STATION	76.53198	9.00066	0	56.37272727	72.85454545	3.687272727	17.12636364

Fig.3. 2011 Data of KMML Station and Literature

4. Methodology

Using SMOTE (Synthetic Minority Over-sampling Technique) with machine learning involves several steps to effectively handle imbalanced datasets. Here's the methodology you can see in Fig 4.

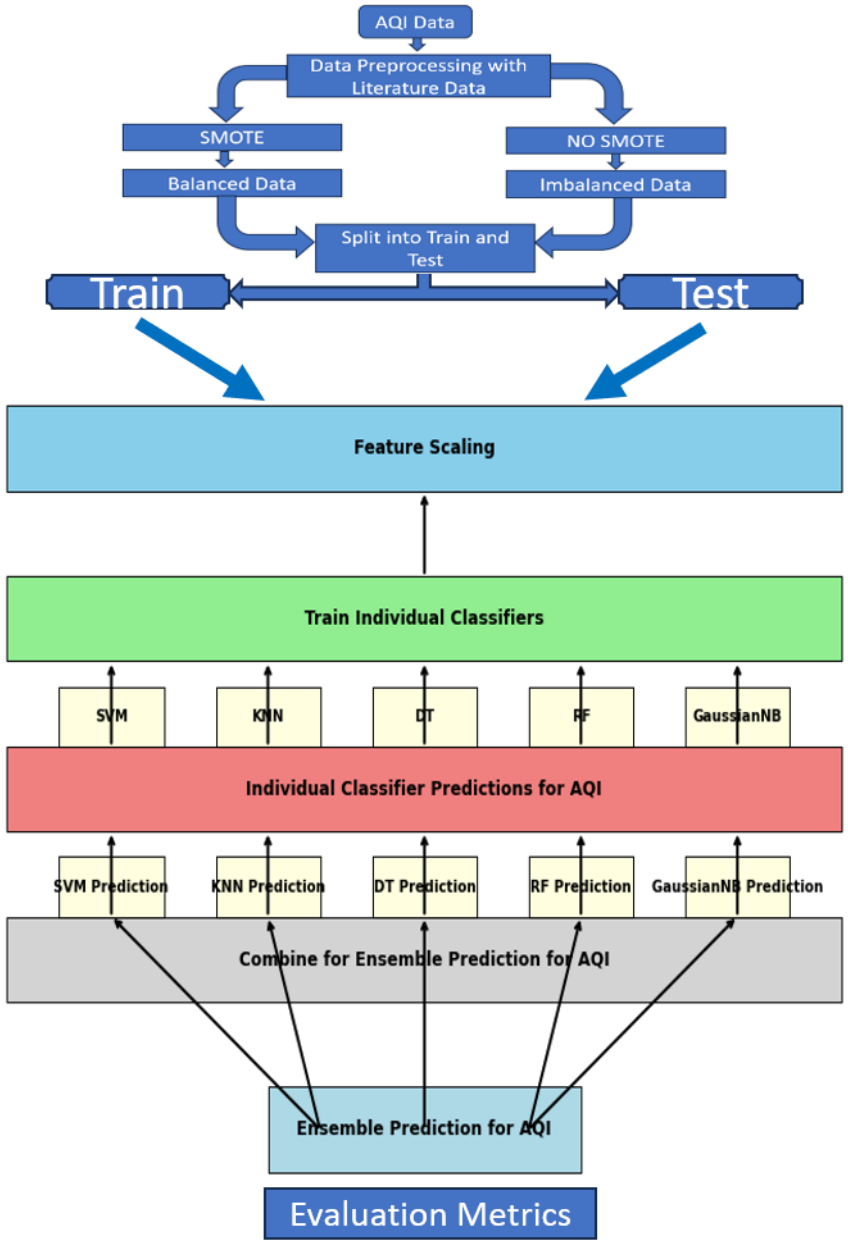

Fig.4. Proposed Flowchart

SMOTE Algorithm

There is asymmetry in the original data set. The procedure for the simulated minority over sampling method (SMOTE) is utilized to rectify the unbalanced data collection. The technique uses over sampling to improve accuracy. To guarantee that every label for each class gets the same amount of rows in a set or the same number of columns as necessary, additional rows are added to the database if necessary. An unbalanced data collection contains asymmetry. A skew range of classes results from an unbalanced dataset, which has several effects on the reliability of the model's predictions.

This means that a balancing of the data is essential. The process of overs using a positive label is one way to increase precision. In this study, oversampling is accomplished with the help of SMOTE. Using a neighbor-based approach, the SMOTE method boosts the representation of underrepresented groups within a dataset.

After that point, it doubles the frequency with which the underrepresented group (the positives) appears, reaching six to twelve times. It helps balance out data sets to boost the efficiency of algorithms while preventing problems with overfitting. Typical implementations of SMOTE involve locating a feature vector as well as its nearest neighbor, computing the

distinction among the two, increasing it by an arbitrary number from zero to one, locating an additional point on the line section after adding the selected number to the vector of features, and so on. SMOTE is preferable to making copies that are slightly off than the original information since it generates completely new data pieces. Fig. 5 depicts the SMOTE model.

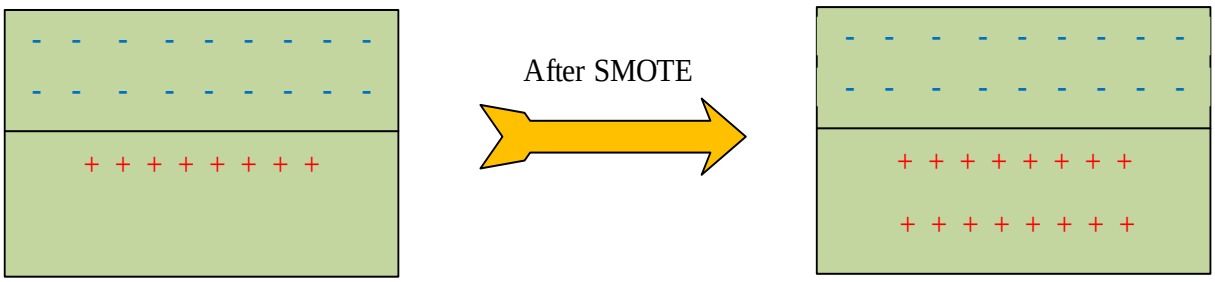

Fig.5 SMOTE Model

In this research, we use the SMOTE method to balance dataset in order to increase model accuracy. A skewed distribution of classes due to an uneven dataset will lead to inaccurate predictions. Balanced datasets lead to more accurate models, more balanced reliability, as well as a better-balanced detection rate. Thus, SMOTE is used to achieve this goal and enhance precision. SMOTE is advantageous since it does not generate identical numbers instead making synthetic data points that deviate from the real data points by a small amount. This approach helps get over the over fitting problem brought on by random sampling by generating instances that are comparable to the minority instances that already exist. Classifiers' generalization skills are enhanced by SMOTE because it generates limits on decisions that are both less specific and more expansive.

Ensemble Machine Learning Classification

In artificial intelligence, a combination framework is a technique where numerous models are combined to provide a superior result than that of just one model. It is the goal of the ensembles predictive framework to either reduce variance (through bagging) or bias (by boosting) or increase prediction quality (via stack). Fig 6 depicts the ensemble machine learning model.

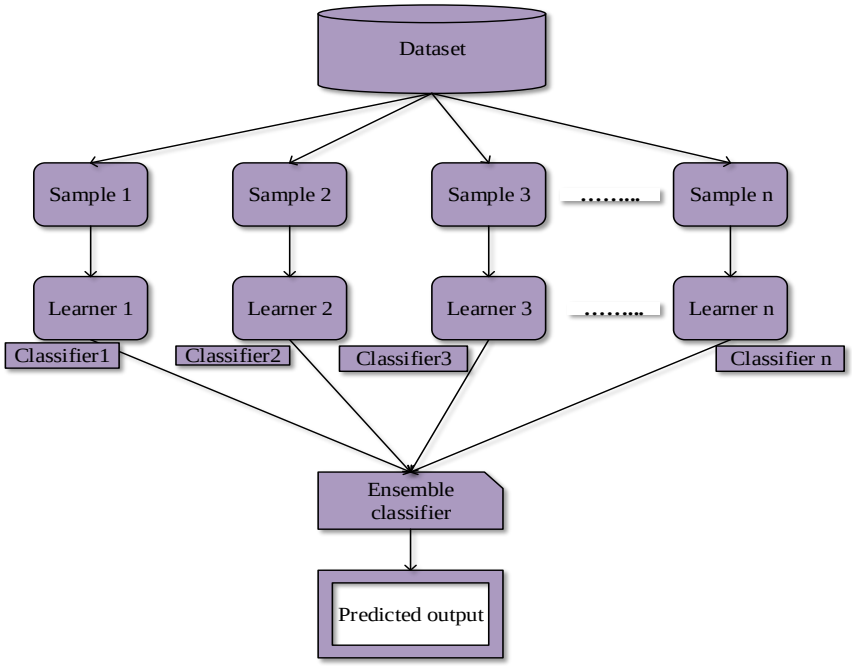

Fig.6 Ensemble Machine Learning Model

SVM

The goal of a support vector machine (SVM) is to form a function with an objective,

then locate a partition hyper plane that may fulfill the class criterion, all while adhering to the notion of reducing structural risk.

(x_i, y_i) = Separable linear dataset.

$i = 1, \ldots, n$, $x \in R^d$ and $y \in \{+1, -1\}$ = Label of class.

$\omega \cdot x + b = 0$ = Partition hyper plane.

Where

ω =Partition of normal vector hyper plane.

b =Offset hyperplane.

A partitioning hyper plane to create the bilateral blank region that is as far away from the location in the training data set as feasible is optimal, i.e., $2/\|\omega\|$, The search for the maximum, which may be described as,

$$\text{Minimize } \phi(\omega) = \frac{1}{2} \| \omega \|^2 \qquad (1)$$

The following is an example of a constraint condition.

$$y_i(\omega \cdot x_i + b) \geq 1 \qquad (2)$$

In this case, the Lagrange function is defined as:

$$L(\omega, b, \alpha) = \frac{1}{2}(\omega \cdot \omega) - \sum_{i=1}^{n} \alpha_i(y_i(\omega \cdot x_i + b) - 1) \qquad (3)$$

Two conditions for the following subject, i.e., $\sum_{i=1}^{n} y_i \alpha_i = 0$ as well as $\alpha_1 \geq 0$, Consequently, the formula below may be used to locate the lagrange function's minimum.

$$\max Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \qquad (4)$$

The best definition of the class function is as follows.

$$f(x) = \text{sgn}\left((\omega^* \cdot x) + b^*\right) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x) + b^*\right) \qquad (5)$$

Nonlinear mapping is used in the case of nonlinear reparability $\kappa(x)$ may be used to transform x into a linear separable feature space in a higher dimension.

a. Decision Tree

The choice tree's base node is constructed first, followed by its child nodes. The knowledge is organized into categories, with the nodes standing in for the defining qualities and attributes that serve as choice points. Nodes are connected to one another at various levels via branches, which stand for the various judgments reached after checking the node's properties. Using a tree data arrangement as well as an if-then expression, it is a guided machine learning approach.

The Desmond Density measure for data abundance is the basis of the choice tree technique. If P is the probability distribution, then $P = (p1, p2, P_3, \ldots, Pn)$, In the equation that follows, where (Pi) is the chance that the number (i) would emerge throughout the process, we can calculate the information contained by this distribution, which we name the Entropy of P with a sample data set S.

$$P = (p1, p2, P3, \ldots, Pn)\ E_{13} \qquad (6)$$

$$\text{Entropy } (P) = \sum_{i=1}^{n} \text{PiXlog (Pi)} \qquad (7)$$

E14

$$\text{Gain } (P, T) = \text{ Entropy } (P) - \sum_{j=1}^{n} (\text{PjX Entropy (Pj)}) \qquad (8)$$

E15

Where Pj is a complete list of all possible values for (T)

KNN

An alternative way to think about the kNN method is as a system of votes, where a fresh data point's classification is decided by the majority of the class labelling of its closest 'k' (in which k is an integers) neighbours in the space of features. Consider a hypothetical election in a tiny town of a few hundred people, in which you must choose a party to represent you. The easiest way to find out which party in politics your neighbors favor is to just ask them. You are more likely to cast a vote for the Republican Party if your 'k' closest neighbors are members of that party. This is analogous to the way the kNN method works, where the information point

with the fewest neighbors is assigned a class identity with the highest probability.

Let's go in further with a different illustration. Just pretend that you have some information regarding fruits like pears as well as grapes. The size and the shape spherical the fruit is also factor towards the final rating. To see this, you choose to create a chart. You can also use this method to identify an unknown fruit by plotting it on the graph as well as calculating its distance to the k (some number) closest spots. Using the coordinate system shown below, I have absolute certainty that we are dealing with a pear since the three closest coordinates all represent pears. Using the four closest points, we can be 75% certain that this is a pear since three of them are pears and one is a grape. Later in this post, we'll discuss the various distance measurement techniques as well as how to get the ideal value for k.

The total amount of neighbours used by the KNN method is defined by the value of k, making it a pivotal variable. In the k-nearest neighbors (k-NN) technique, the quantity of k must be determined by analyzing the data at hand. A greater number of k is preferable if the provided data is more prone to anomalies or noise. To prevent categorization draws, it is preferable to use an odd number for k. To determine the optimal k for a certain data set, cross-valid techniques may be used.

Random Forest

The Random Forest algorithm is a complex ensemble method that involves multiple decision trees. Here, the key mathematical components and operations involved in the Random Forest algorithm is outlined in the Table 1.

Training Process:

Bootstrap Sampling:

- Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the training dataset with N samples, where $x_i \in \mathbb{R}^d$ represents the input features and $y_i \in \mathcal{Y}$ denotes the corresponding label (for classification, $\mathcal{Y}$ is the set of class labels; for regression, $\mathcal{Y} = \mathbb{R}$).
- Random Forest begins by creating B bootstrap samples D_b from D, where each bootstrap sample D_b is created by sampling N examples with replacement from D.

Decision Tree Construction:

- For each bootstrap sample D_b:
  - Randomly select a subset of features $F_b \subseteq \{1, 2, \ldots, d\}$ with $|F_b| \ll d$.
  - Construct a decision tree T_b using D_b and F_b:
    - The tree recursively splits nodes to minimize impurity (e.g., Gini impurity for classification, variance reduction for regression) until a stopping criterion is met (e.g., maximum depth, minimum samples per leaf).

Prediction Process:

Aggregation:

- After training B decision trees $\{T_b\}_{b=1}^{B}$:
  - **Classification**: For a new input x, predict its class label $\hat{y}(x)$ using majority voting:

$$\hat{y}(x) = \text{mode}(\{T_b(x)\}_{b=1}^{B})$$

    where $T_b(x)$ is the prediction of tree T_b for input x.
  - **Regression**: For a new input x, predict its output $\hat{y}(x)$ using the average prediction:

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

    where $T_b(x)$ is the prediction of tree T_b for input x.

Feature Importance:

Importance Calculation:

- Random Forest can also calculate feature importance, which indicates the contribution of each feature j in reducing impurity across all trees:
  - **Mean Decrease Impurity**: Calculate the average decrease in impurity (e.g., Gini impurity) caused by splits over feature j across all trees.
  - **Mean Decrease Accuracy**: Permute the values of feature j in the out-of-bag samples and measure the decrease in accuracy, then average over all trees.

The Random Forest algorithm combines the power of bootstrap sampling, decision tree construction with feature randomness, and ensemble aggregation to provide robust predictions and feature importance insights.

Gaussian Naive Bayes (GaussianNB)

Gaussian Naive Bayes (GaussianNB) is a variant of the Naive Bayes algorithm that is well-suited for classification tasks where the features are continuous and assumed to follow a Gaussian (normal) distribution. Here's an explanation of GaussianNB:

Naive Bayes Algorithm: Naive Bayes is a probabilistic classifier based on Bayes' theorem, which assumes independence between features given the class. Despite its simplistic assumptions, Naive Bayes often performs well in practice and is efficient for large datasets.

GaussianNB Specifics: Feature Assumption: GaussianNB assumes that the continuous features follow a Gaussian distribution. This means that for each class, the distribution of each feature is estimated using the mean and variance of the observed values in the training data. The classification process is depicted in the Table 2.

- Given a set of features $X = (x_1, x_2, ..., x_d)$, GaussianNB calculates the conditional probability of each class C_k given the features using Bayes' theorem:

$$P(C_k|X) = \frac{P(X|C_k) \cdot P(C_k)}{P(X)}$$

where:

- $P(C_k|X)$ is the posterior probability of class C_k given features X.
- $P(X|C_k)$ is the likelihood of observing X given class C_k.
- $P(C_k)$ is the prior probability of class C_k.
- $P(X)$ is the evidence, which acts as a normalizing constant.

Training and Prediction: During training, GaussianNB estimates the mean and variance of each feature for each class from the training data. During prediction, it computes the probability of each class given the input features and selects the class with the highest probability as the predicted class.

Classification

Computation of Air Quality Index: The five levels of the established Indian the air quality index reflects varying degrees of air pollution:

Good: The AQI for a given area might range from zero to one hundred. There is minimal to no danger from pollution in the air, as well as the air condition is good.

Moderate: Among 101 to 200, that's a community's AQI. The air is generally safe to breathe, although a tiny proportion of individuals may have some mild health effects from a few specific contaminants. Some persons may be extremely susceptible to ozone and have respiratory problems as a result. Sensitive populations might experience health impacts with AQI levels among 201 and 300.

Poor: This suggests that they will have milder effects than the overall population. For instance, exposure to ozone poses a larger threat to persons who already have lung illness, and particle pollution poses a bigger threat to those who already have respiratory illnesses or heart disease. Whenever the AQI falls within this range, there is little risk of harm to the general population.

Very Poor: When the Air Quality Index (AQI) ranges from 301 as well as 400, it is possible for everybody to feel the consequences on their health. Health impacts may be more severe for people in vulnerable populations.

Severe: A health warning is issued for the whole population if the value is between 401 and 500. The following Eqs may be used to determine the individual sub-indices that make up the IND-AQI. (1) and (2)

q=100(V/Vs) (9)

Where,

q=Rating Quality;

V= Values of observed parameters and

V" " S= Recommended value of standard parameter (CPCB, MoEF, 1998, 2009).

If' 'n' parameters are taken into account, the Air Quality Index (AQI) is calculated as the Arithmetic Mean of these 'n' Quality Ratings.

Geometric mean, g= anti log{(log a+log b+⋯log x)/n

Where a, b, c, d, x= air quality ratings of varying levels; and n= rating scale for air

quality n.

The sum of all the sub-AQIs is presented as the IND-AQI, the overall AQI.

Oversampling, particularly using techniques like SMOTE, involves creating synthetic samples for the minority class by interpolating between existing samples. This helps balance class distribution and prevents the model from being biased towards the majority class, improving its ability to generalize and make accurate predictions for the minority class.

5. Results & Discussion

Training and evaluation sets will be created when the dataset is cleaned, reduced, and set up to our specifications. To allow for the methods to be readily applicable in a real-world use case, the aim is to employ the simplest, most simplistic implementations possible. To determine which of these three techniques is more precise, we'll use a variety of criteria to make comparisons among them. By comparing several approaches to predicting the AQI, we may learn more about our options as well as choose the one that best fits our needs. We will also use the SMOTE method to compare the precision levels achieved using an uneven as well as an even database. Therefore, the technique is an ordered procedure, with the first stage being the identification and preparation of an appropriate dataset. After that, SMOTE is used for some more data preparation to get the dataset closer to parity. In order to highlight any modifications for efficiency that may develop owing to weighing, both balanced and unbalanced datasets will be retained and utilized. After that, the training set is often divided into training sets in order to train the predictions and assess their accuracy against actual data, as is customary in automated learning procedures. Normalization as well as scaling of features are performed. Fig 7 shows the concentration distribution graphs for the pollutants after pre-processing.

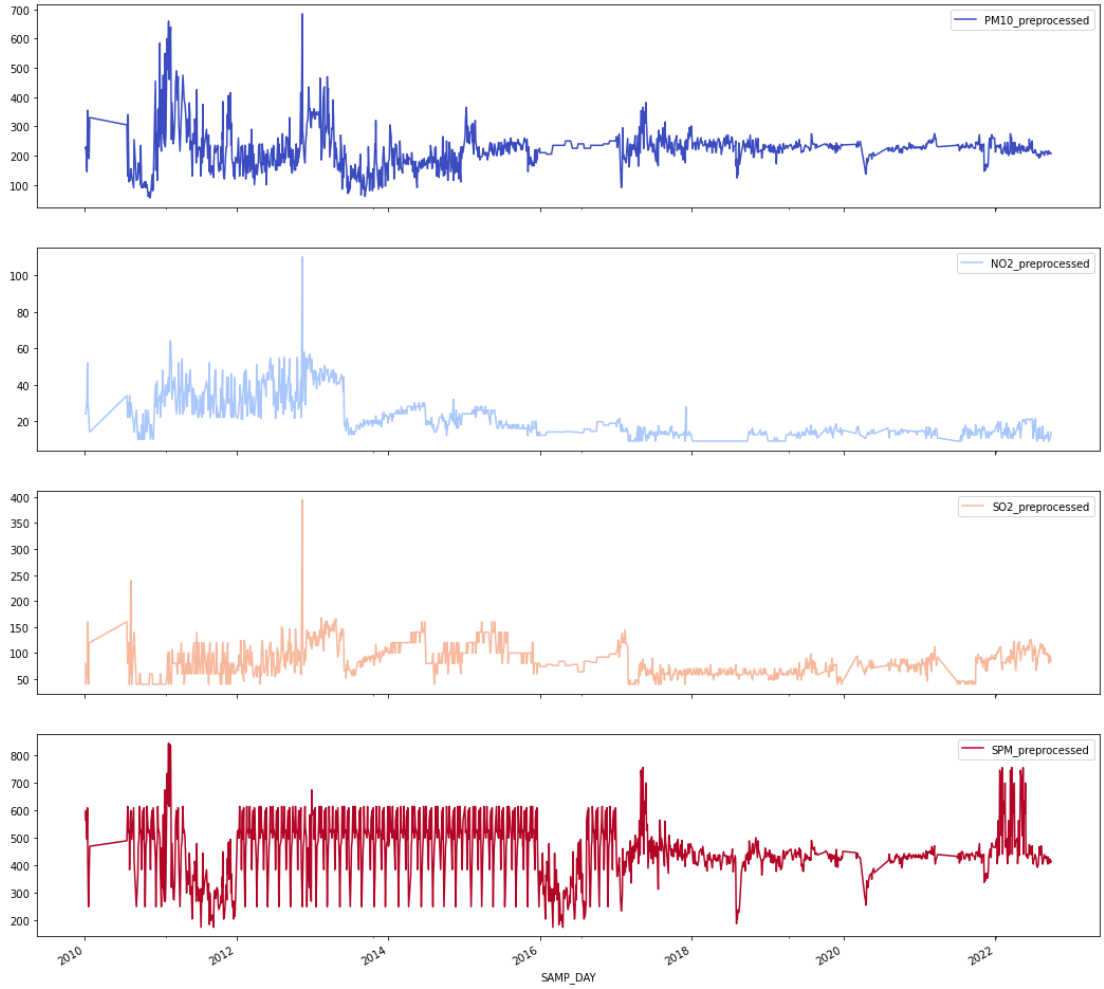

Fig 7. The concentration distribution graphs of the trended Pollutant data

We met our needs by the trended data of Chavara surroundings from 2010 to 2022. Data preprocessing

The data for each pollutant in the Chavara monitoring station for the years 2010 to 2022 were obtained by a linear trend determined by the closest known values from the literature. It is cleaned up in data processing. It is crucial to examine the levels of pollution in the KMML surroundings due to the fact that it is the primary source of pollutants in the subject area. The data sets were cleaned by deleting rows with values that were null. Unneeded, insignificant, or incorrect information is purged in Anaconda Python. The category imbalance in the AQI_Bucket results is corrected using a synthetic minority over sampling method (SMOTE) once the dataset has been cleaned. It took 3 manual iterations, to establish a good degree of balance of all AQI categories, like Good, Moderate, Poor, Very Poor, and Severe. This is done so that the final dataset is more representative of most industrial areas. For, without SMOTE, the artificial minority over sampling method (SMOTE) is has not utilized right after the dataset has been cleaned of extraneous, irrelevant, and incorrect information. A ratio of 80:20 separates the datasets into train and test data. These will be put to use in both the training and testing phases of the model's development. Machine learning systems' predictions are validated against the raw data to improve accuracy. Using 80 percent of the information to train the model with has been shown to be optimal in empirical studies. To create the test and instruction sets, the data is sampled at random. It has achieved widespread popularity and acceptance. The best outcomes may be achieved, according to empirical research, by using the remaining twenty percent of the information for the purpose of testing. To create the test as well as training sets, the data is sampled at random. It has achieved widespread popularity and acceptance. In an effort to render the information usable as well as consistent, the information has been standardized. The Scikit-Learn Library's Standard Scaler was utilized for this purpose. The characteristics are normalized by dropping the mean and adjusting the variance down to the unit level. Different algorithms, including KNN, SVM, DT, RF, and GaussianNB analysis, are utilized to predict the index of air quality after standardizing the variety of characteristics in the data sets. These methods are subsequently contrasted to determine which one provides the greatest precision level. Then using voting method, the prediction in ensemble has been done. The AQI quality for all pollutants is assessed with the use of artificial intelligence algorithms. The levels of precision are shown graphically. Calculation of evaluation metric for each ML technique KNN, SVM, DT, RF, GaussianNB, Ensemble, and Ensemble SMOTE are assessed using accuracy, precision, recall and F1-score.

1334x12 table

	1 SAMP_DAY	2 MONITOR_STN	3 SO2_preprocesse	4 NO2_preprocessed	5 PM10_preprocessed	6 SPM_preprocessed	7 PM10_SubIndex	8 SO2_SubIndex	9 NO2_SubInd	10 SPM_SubIndex	11 AQI	12 AQI_BUCKET
1	NaT	'Chavara'	80	24	225	565	183.3333	100	30	417.8571	418	3
2	NaT	'Chavara'	40	24	230	600	186.6667	50	30	442.8571	443	3
3	NaT	'Chavara'	60	32	145	495	130	75	40	367.8571	368	4
4	01/13/2010	'Chavara'	60	52	285	610	235	75	65	450	450	3
5	01/15/2010	'Chavara'	160	30	355	420	306.2500	126.6667	37.5000	314.2857	314	4
6	01/18/2010	'Chavara'	40	26	250	250	200	50	32.5000	150	200	1
7	01/21/2010	'Chavara'	120	18	190	360	160	113.3333	22.5000	366.6667	367	4
8	01/24/2010	'Chavara'	120	14	330	470	280	113.3333	17.5000	350	350	4
9	07/22/2010	'Chavara'	160	34	305	490	255	126.6667	42.5000	364.2857	364	4
10	07/26/2010	'Chavara'	80	22	130	615	120	100	27.5000	453.5714	454	3
11	07/28/2010	'Chavara'	80	30	340	520	290	100	37.5000	385.7143	386	4
12	NaT	'Chavara'	120	26	110	530	106.6667	113.3333	32.5000	392.8571	393	4
13	NaT	'Chavara'	40	32	130	500	120	50	40	371.4286	371	4
14	NaT	'Chavara'	60	34	115	385	110	75	42.5000	408.3333	408	3
15	NaT	'Chavara'	60	22	115	565	110	75	27.5000	417.8571	418	3
16	NaT	'Chavara'	240	30	155	600	136.6667	153.3333	37.5000	442.8571	443	3
17	08/15/2010	'Chavara'	40	26	130	495	120	50	32.5000	367.8571	368	4
18	08/24/2010	'Chavara'	120	18	90	610	90	113.3333	22.5000	450	450	3
19	08/26/2010	'Chavara'	140	14	255	420	205	120	17.5000	314.2857	314	4

Fig.8. Input Dataset

Fig 8 shows the sample values of the input dataset. In this, the chavara dataset is used for predicting the Air Quality Index values. The dataset comprises of 12 variables and 1334 instances as shown in the above figure. The first two variables denote the date and place of the recording. The next 8 variables show the chemical component in that place at that time. The 11th variable gives the Air Quality index value. The final 12th variable denotes the AQI level like 'Good' :0, 'Satisfactory' :1, 'Moderate' :2, 'Poor' :3, 'Very Poor': 4, 'Severe': 5.

The first stage of pre-processing is performed by removing the first two variables, as it has lower impact in AQI prediction. These two variables removed and is then used for further processing. Fig 9 shows the data after first stage pre-processing. Fig 10 depicts the correlation of all the pollutant concentrations against AQI level.

_preprocessed	NO2_preprocessed	PM10_preprocessed	SPM_preprocessed	PM10_SubIndex	SO2_SubIndex	NO2_SubIndex	SPM_SubIndex	AQI	AQI_BUCK
80	24	225	565	183.33	100	30	417.86	418	3
40	24	230	600	186.67	50	30	442.86	443	3
60	32	145	495	130	75	40	367.86	368	4
60	52	285	610	235	75	65	450	450	3
160	30	355	420	306.25	126.67	37.5	314.29	314	4

Fig 9. First Stage pre-processed

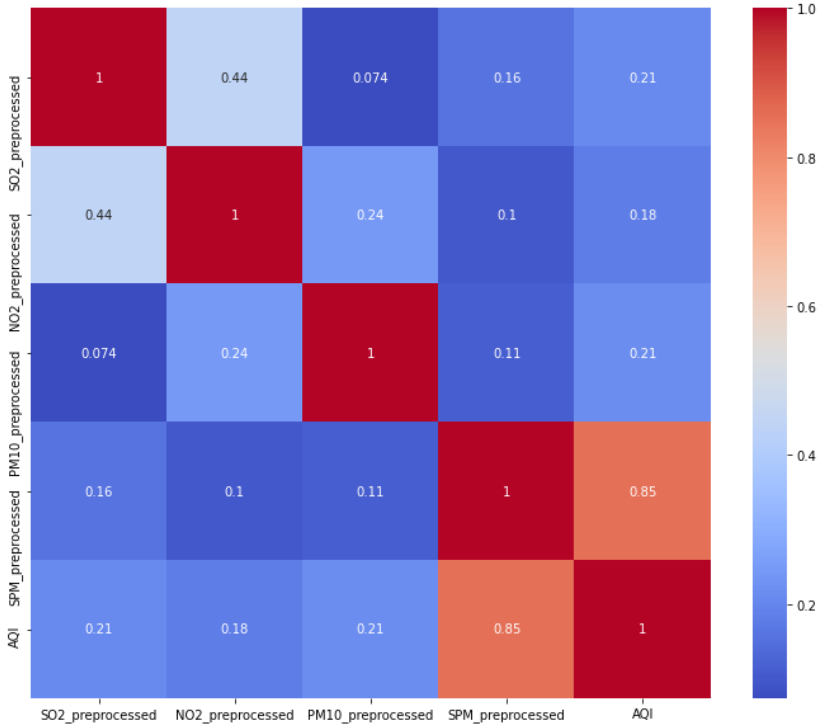

Fig 10. Correlation of all Pollutants against AQI level

The first stage pre-process is followed by the second stage pre-processing steps called SMOTE to balance the data. As in original data, the number of samples in each class are not equally distributed as shown in the below Fig 11. Then, these data are processed using SMOTE to evenly distribute the classes as shown in the below Fig 12.

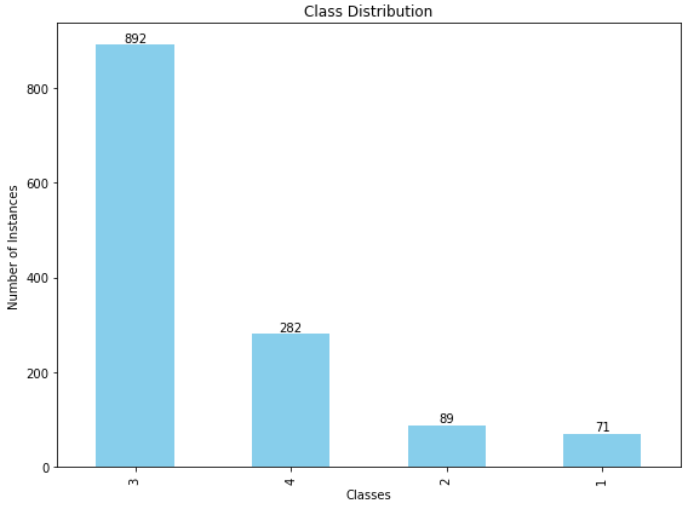

Fig 11. Before SMOTE processing

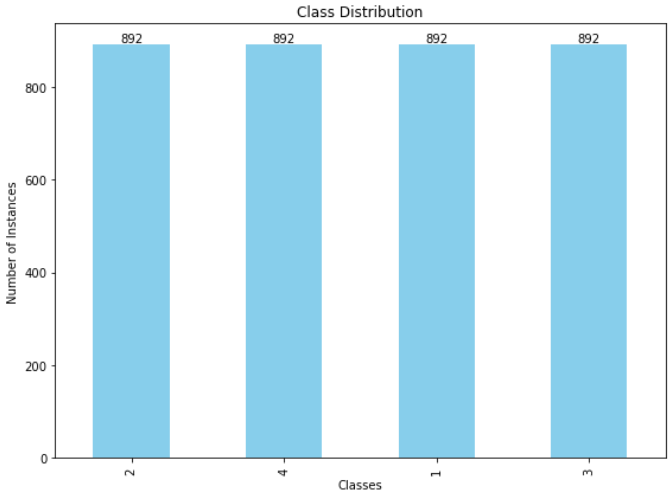

Fig 12. SMOTE data

Fig 12 shows that the number of samples in each class is evenly distributed and the number of samples in the set is also increased to 3568. This increased dataset is then split into training and testing using hold-out approach with 80% as training and 20 % testing. Then, the individual models like KNN, SVM, DT, RF, and GaussianNB were trained using 80% of data and then tested with 20% of data. Fig 13 depicts the evaluation metrics of the classifiers.

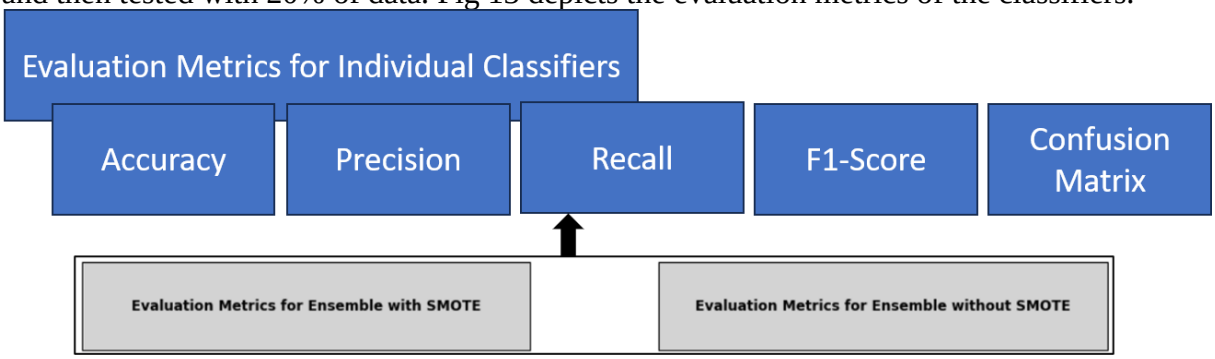

Fig 13 Evaluation Metrics

From Fig 14, It is observed that the GaussianNB model accuracy is minimum, 94.51 as compared to other models. This has been enhanced by combining all the model result using voting system and the corresponding accuracy for the Ensemble model is 99.75. After data balancing has been done using SMOTE, the individual classifiers' as well as the combined model's accuracy has considerably increased. Hence, Ensemble SMOTE model gives the highest accuracy of 100.

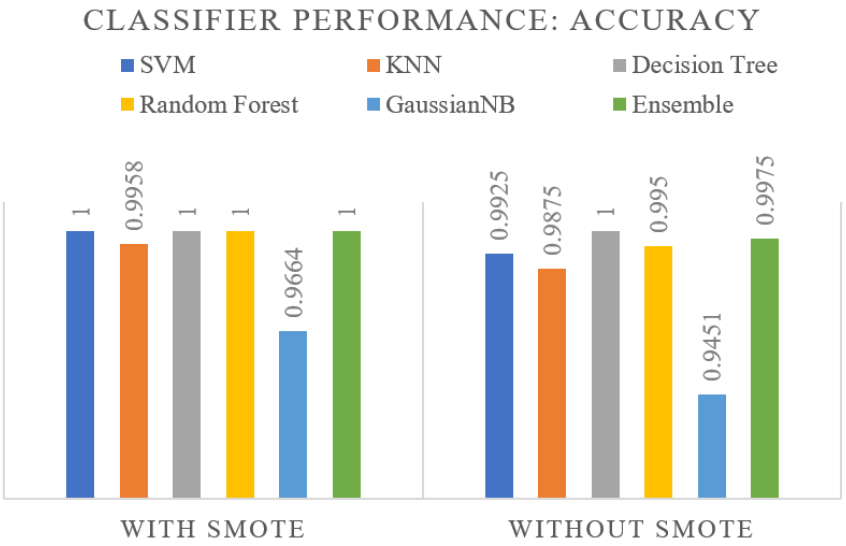

Fig 14. Comparative Accuracy Analysis of Classifiers

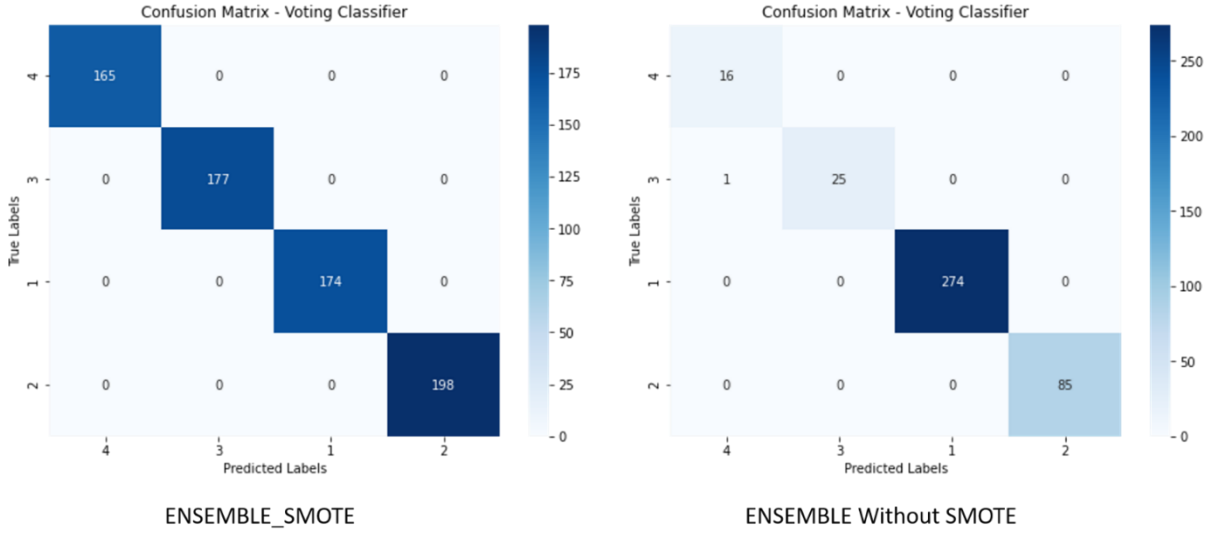

Fig 15. Ensemble model Confusion Matrix.

From the Fig 15, it can be observed that the proposed Ensemble SMOTE model can achieve the higher accuracy as compared to Ensemble model. Based on this, the proposed Ensemble SMOTE method's performance is compared with the KNN, SVM, DT, RF, GaussianNB, and Ensemble models in terms of Precision, recall, and F1-Score.

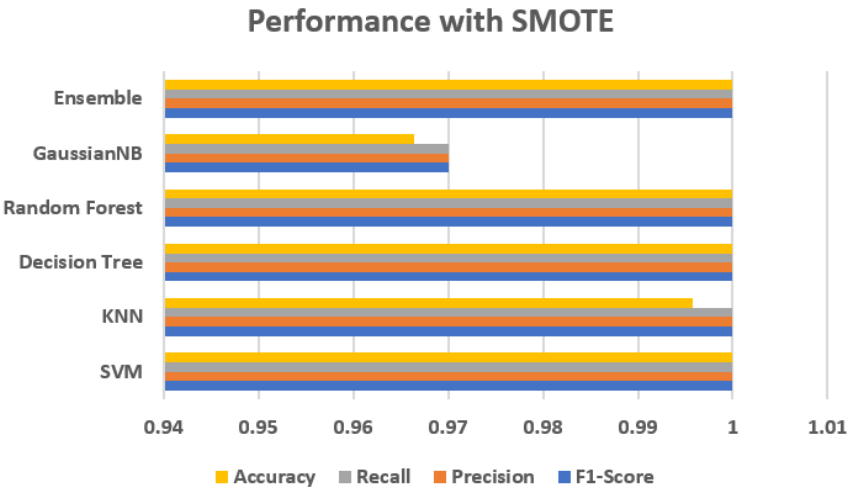

Fig 16. Performance with SMOTE.

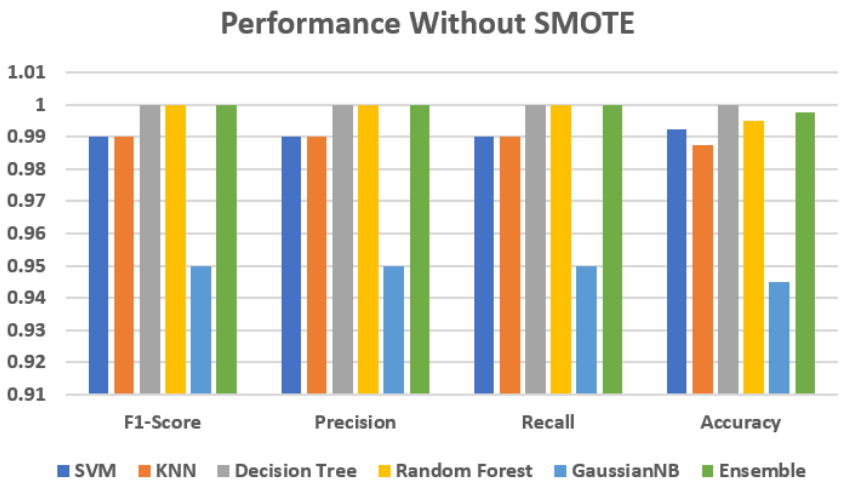

Fig 17. Performance Without SMOTE.

Fig 16 and Fig 17 shows the comparison charts of the proposed Ensemble SMOTE approach with the individual classifiers and the ensemble classifier. From the results, it can be observed that the proposed approach achieved better results as compared to the existing and individual classifiers. The study employed machine learning algorithms and SMOTE to predict Air Quality Index (AQI) levels in the study area. The ensemble SMOTE method outperformed individual classifiers like KNN, SVM, DT, RF, GaussianNB, and ensemble achieving higher accuracy, precision, recall, and F1-score, indicating its effectiveness in predicting AQI levels. The study also highlighted the importance of data preprocessing and balancing for improved prediction accuracy.

6. Conclusion

With the advent of Industry 4.0, conventional manufacturing is being transformed into smart manufacturing, opening up new possibilities, and allowing machines to comprehend processes, communicate with their surroundings, and intelligently adjust their behavior. Real-time air quality monitoring and assessment is desired for Industry 4.0 due to the development of IoT infrastructures and machine learning methods. AI's machine learning (ML) branch has emerged as the primary force behind these developments in the industrial sectors, offering the chance to improve decision-making and expedite discovery procedures even more. On the other hand, machine learning algorithms (ML) get their knowledge directly from data, examples, and experience, and then use this knowledge to generalize to solve complex problems. In order to provide solutions to such questions and prevent future concerns, this study highlights the challenges and future developments of machine learning applications for air quality assessment in the KMML Industrial Area (KNN, SVM, DT, RF, GaussianNB). In future, Data Augmentation model is proposed for addressing data scarcity issues and also the deep learning model is used for improving the accuracy.

Nomenclature

SMOTE	Synthetic Minority Over-sampling Technique
SO2	Sulfur Dioxide
NO2	Nitrogen Oxides
AQI	Air Quality Index
PM10	Particulate Matter 10 micrometers or less in diameter
SPM	Total Suspended Particulate Matter

Declaration:

Ethics Approval and Consent to Participate:

No participation of humans takes place in this implementation process

Human and Animal Rights:
No violation of Human and Animal Rights is involved.
Funding:
No funding is involved in this work.
Data availability statement:
Data sharing not applicable to this article as no datasets were generated or analyzed during the current study
Conflict of Interest:
Conflict of Interest is not applicable in this work.
Authorship contributions:
All authors are contributed equally to this work
Acknowledgement:
There is no acknowledgement involved in this work.

References

1. Farihah Mohmad Shamsuddin, A., Jalaludin, J., & Haslina Hashim, N. (2022). Exposure to industrial air pollution and its association with respiratory symptoms among children in Parit Raja, Batu Pahat. IOP Conference Series: Earth and Environmental Science, 1013.
2. Vozdova, M., Kubíčková, S., Kopecká, V., Šípek, J., & Rubes, J. (2022). Association between sperm mitochondrial DNA copy number and deletion rate and industrial air pollution dynamics. Scientific Reports, 12.
3. Han, X., Dou, J., & Tang, C. (2022). Polycentricity, Agglomeration, and Industrial Air Pollution in the Chinese City-Regions. Frontiers in Environmental Science.
4. Han, C., Hua, D., & Li, J. (2023). A View of Industrial Agglomeration, Air Pollution and Economic Sustainability from Spatial Econometric Analysis of 273 Cities in China. Sustainability.
5. Dimitrova, M., Trenchev, P., Avetisyan, D., & Spasova, T. (2023). Spatio-temporal monitoring of air pollution over Bulgaria's largest industrial area using Sentinel-5p TROPOMI data. International Conference on Remote Sensing and Geoinformation of Environment.
6. Liu, Y., Xie, C., Wang, Z., Rebai, N.H., & Lai, X. (2023). The Role of Industrial Structure Upgrading in Moderating the Impact of Environmental Regulation on Air Pollution: Evidence from China. Atmosphere.
7. Qi, G.Z., Wang, Z., Wang, Z., & Wei, L.R. (2022). Has Industrial Upgrading Improved Air Pollution?—Evidence from China's Digital Economy. Sustainability.
8. Le, V., Nguyen, D.H., Wang, H., Liu, B., & Chu, S. (2022). Efficient UAV Scheduling for Air Pollution Source Detection From Chimneys in an Industrial Area. IEEE Sensors Journal, 22, 19983-19994.
9. Vozdova, M., Kubíčková, S., Kopecká, V., Šípek, J., & Rubes, J. (2022). Effects of the air pollution dynamics on semen quality and sperm DNA methylation in men living in urban industrial agglomeration. Environmental and Molecular Mutagenesis, 63, 76 - 83.
10. Połednik, B. (2022). Emissions of Air Pollution in Industrial and Rural Region in Poland and Health Impacts. Journal of Ecological Engineering.
11. Devasekhar, M.V., & Natarajan, D.P. (2023). Prediction of Air Quality and Pollution using Statistical Methods and Machine Learning Techniques. International Journal of Advanced Computer Science and Applications.

12. Zhang, X., Sun, Z., Zhou, Z., Jamali, S., & Liu, Y. (2022). Analysis and Dynamic Monitoring of Indoor Air Quality Based on Laser-Induced Breakdown Spectroscopy and Machine Learning. Chemosensors.
13. Gupta, N.S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. Journal of Environmental and Public Health.
14. G.Kalaivani, Scholar, R., & Kamalakkannan, S. (2022). Web Scraping Technique for Prediction of Air Quality through Comparative Analysis of Machine Learning and Deep Learning Algorithm. 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 263-273.
15. Chakravarty, A., S, S.S., & S, S. (2022). An Exploratory Analysis of Delhi Air Quality Using Statistics and Machine Learning Models. 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 1-6.
16. Liu, Q., Cui, B., & Liu, Z. (2024). Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling. Atmosphere.
17. Sonawane, P., Dhanawade, S., Barangule, V., Kulkarni, A., & Mahalle, P.N. (2023). Air Quality Analysis & Prediction Using Machine Learning: Pune Smart City Case Study. 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), 1-6.
18. Hardini, M., Riza Chakim, M.H., Magdalena, L., Kenta, H., Rafika, A.S., & Julianingsih, D. (2023). Image-based Air Quality Prediction using Convolutional Neural Networks and Machine Learning. Aptisi Transactions on Technopreneurship (ATT).
19. Gowda, D., Kumar, G.S., Sriharsh, B., Chandan, S., & Sushmitha, M. (2023). MACHINE LEARNING APPROACH FOR PREDICTING AND ANALYZING AIR QUALITY.
20. Abdulganiyu, A.O., Kolo, J.G., & Usman, A.U. (2023). DEVELOPMENT OF AN INTERNET OF THINGS BASED AIR QUALITY MONITORING SYSTEM USING MACHINE LEARNING. International Journal of Advanced Natural Sciences and Engineering Researches.
21. D.S., Jaya. (2014). Air Quality Assessment in the Surroundings of KMML Industrial Area, Chavara in Kerala, South India. Aerosol and Air Quality Research. 14. 10.4209/aaqr.2013.10.0327.
22. Lars Barregard, Erik Holmberg, Gerd Sallsten, Leukaemia incidence in people living close to an oil refinery, Environmental Research, Volume 109, Issue 8, 2009, Pages 985-990, ISSN 0013-9351, https://doi.org/10.1016/j.envres.2009.09.001.
23. Machaczka, O., Jiřík, V., Janulková, T. et al. Comparisons of lifetime exposures between differently polluted areas and years of life lost due to all-cause mortality attributable to air pollution. Environ Sci Eur 35, 73 (2023). https://doi.org/10.1186/s12302-023-00778-5