Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7546452

ADVANCE AIR POLLUTION PREDICTION IN OMAN THROUGH MACHINE LEARNING APPROACHES: ANALYSIS, MODELS, AND FUTURE DIRECTIONS

[1]Saif Nasser Al Rumhi,
M.Sc Data Science
Global College of Engineering and Technology
202221125@gcet.edu.om

Corresponding Author
[2]Dr.Janaki Sivakumar
Associate Professor
Global College of Engineering and Technology
janaki.s@gcet.edu.om

Abstract

The study investigates the use of machine learning for air pollution prediction in Oman using both regression and classification models. Data pre-processing and feature engineering are heavily employed so that the data is of quality and is used properly. Moreover, data imbalances, sparsity, and missing values are addressed. Different regression models including Linear Regression, Huber Regression, MLP and classification models such as Logistic Regression and Random Forest are implemented. Yet, the proposed machine learning model including Linear Regression, Random Forest, and Gradient Boosting Regression appear to be the most powerful. Simultaneously, it is also imperative to point out various limitations such as overfitting, the complexity of the model, and the availability of data. There are ways in which the proposed model can be improved, such as potential extension of data collection, location-specific real-time monitoring solutions complemented with different socio-economic indicators. Moreover, using ensemble learning can increase classification accuracy and can produce more interpretable results. The findings can be adopted to evidence-based decision-making and improve air pollution management and public health in Oman.

1. Introduction

In recent years, the extremely rapid industrialization in line with economic growth has substantially intensified environmental pollution levels [1]. This increasing problem predominantly caused by industrial activities is a very serious risk to human health and to the balance of the biosphere. Among the various forms of pollution, air pollution is notable because its effects on health and the climate are immediate, and widespread [2]. Air pollution is marked by the presence in the air of harmful gases, solid particles, liquid droplets, and biological materials such as pollen or spores. These pollutants fall into two categories: primary and secondary.

Air pollution is not caused by just one thing but every direction of fire-truck smoke is inimical to good health. The problem increases seasonally as well as at other times of year. Sandstorms and other natural anomalies--such as burning agricultural waste for fuel--also make significant

contributions to air pollution in certain locales. This complex mesh of sources spews an entire gamut of gases into the atmosphere: Farm chemicals... Industrial, municipal, and even wood smoke dioxins. Carbon monoxide (CO) along with some nitrous gases and sulfurous gases in which there are few, if any benefits other than to warm the atmosphere. Air pollution has taken a terrible toll on people's lungs and hearts: about 1.1 billion people breathe unhealthy air with a death toll of seven million per year worldwide. The health impacts are uneven, with developing countries, mostly in Asia but also including Africa, bearing the brunt of the deaths and illnesses caused by air pollution in all its myriad forms. This dilemma is evident to most clearly seen in the industrial growth and large amounts of pollution discharged by developing countries like India.

The link between air pollution and human disease is profound. A study shows strong evidence that asthma and other lung diseases are connected with atmospheric pollutants [6]. The WHO implications of this link are significant. The WHO has developed guidelines for the regulation of certain unsafe gases. The chief among them are ozone (O3), nitrogen dioxide (NO2) and sulfur dioxide (SO2). However, there are fails in the measurement and control of air pollution. This is because infrastructure lacks mobility[1], and operational complexities are real. Traditional Air Quality Monitoring (AQM) stations have the added problem of high maintenance costs.

As a key member of the Gulf Cooperation Council (GCC), Oman is dealing with the environmental problems that most rapidly developing economies do. Among them is the worst of all: air pollution. The Sultanate has seen a rise in industrial activity and urbanization. Emissions that have shot up steeply are the main source of the pollutants. Sultanate's weather predisposes it to frequent dust storms which only make the air quality worse [6].

To address this, the Omani government has taken several initiatives to deal with air pollution. Among them, environmental regulations that are designed to reduce industrial emissions and raise public awareness. The Ministry of Environment and Climate Affairs monitors air quality closely and works to control pollutants [7]. Furthermore, Oman's Vision 2040 stresses that the country will adhere to those practices that are found in international norms and consider the development of green technologies and renewable energy as being important. This vision is aligned with the world's top-level environment standards [8].

Here we aim to employ advanced data analytics and machine learning methods to study air pollution trends as shown in Figure 1 . The analysis will focus on identifying primary sources of air pollution, evaluating the effectiveness of current mitigations, and projecting future air quality problems under scenarios. Industrial emissions, vehicular traffic, and meteorological conditions, as well as factors such as topography will all be considered in the analysis [13]. Given Oman's strategic economic shift and advanced technology integration into environmental management, this study is both timely and relevant. It aims to provide guidance to policymakers and to introduce more effective methods for air quality management in Oman [12].

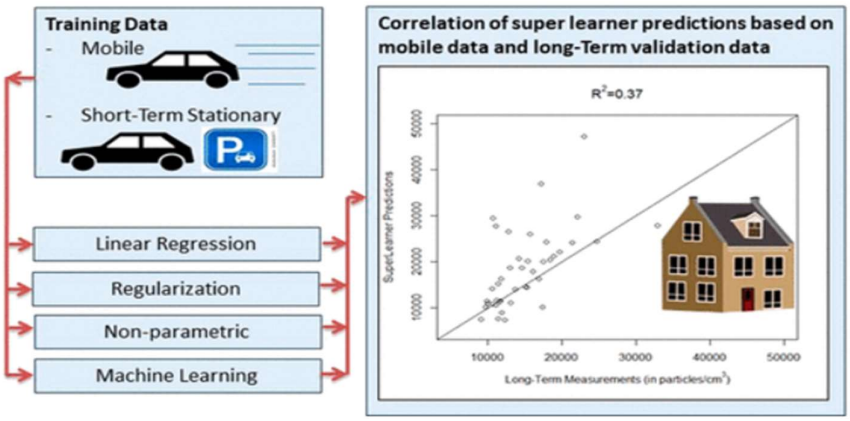

Figure 1. Working of Machine learning models

1.1 Problem Statement

During the winter in Muscat, there was a huge dust and sand cloud. This unexpected natural phenomenon concerned authorities for its effects on the air which seemed to differ from what was usual for the time. Anthropogenic activities are usually lower during the winter months. Intuitively, one might expect better air quality. The sudden appearance of this cloud, however, suggested unexplained or unconventional coal mines, too rigorous industry, or heavens knew what strange conditions that could form smog lumps from pollutants left over by other dust storms, disasters with large numbers killed. Intrigued by the strange-looking dust cloud that appeared over the Muscat air, I called the Environmental Authority hotline to inquire whether these were isolated events or signs of more deep-seated environmental problems.

Not previously mentioned are a few depositional events that occurred over the mountainous regions. This interaction with the Environmental Authority showed that air quality management is a complex process. Further, it suggested potential deficits in the monitoring system, which was based on general models or calculations that did not take account of something that might occasionally happen.

Given the current explosion in machine learning technology, there is a chance to create an air quality analysis model that is more precise, versatile--indeed several steps from being state-of-the-art in its field. Here are some other important factors that need to be considered:

1.2 Purpose and objective of the study

This research looks into which method of regression is most applicable for predicting air pollution levels in Oman. The application of machine learning techniques to environmental science is not only helping the advancement of this field but also benefiting environmental protection and public health.The objectives of this study were:

1. **Environmental scientific research:** The investigation contributes to the field of environmental science by applying sophisticated regression techniques in order to solve the problem of pollution in the air.
2. **Environmental protection:** One of the objectives of the study is to provide insights that can aid in environmental protection efforts.
3. **Public health:** Early, accurate predictions can enable policymakers and health care professionals to take appropriate measures aimed at reducing the adverse health effects .

The Air Quality Index (AQI) is obtained by applying the correlation coefficient (r) to evaluate the degree of pollution in an area as shown in Table 1.

Table1. Air Quality Index(AQI)[19]

Quality ex (AQI) Values	**Level of Health Concern**	**Colors**	**Meaning**
0 to 50	Good	Green	Air quality is considered to be satisfactory; pollution poses little or no risk.
51 to 100	Moderate	Yellow	Air quality is acceptable: however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
101 to 150	Unhealthy for Sensitive Groups	Orange	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
151 to 200	Unhealthy	Red	Members of sensitive groups may experience more serious health effects.
201 to 300	Very Unhealthy	Purple	Health alert: everyone may experience serious health effects.
301 to 500	Hazardous	Maroon	Health warning of emergency conditions. The entire population is more likely to be affected.

2. Literature Review

Table 2. Few experiments form Literature

Serial Number	**Title**	**Author and year**	**Data Set Used**	**Demography**	**Parameters Used**
S1	Air Pollution Analysis Using Enhanced K-Means Clustering Algorithm for Real	Kingsy Grace, R. I Manimegalai, R. Geetha Devasena, M. S Rajathi, S. Usha, K. Raabiathul Baseria, 2016.	RealTime dataset details not mentioned.	urban areas where air pollution monitoring is critical	The study involved air pollution parameters like sulfur dioxide, nitrogen dioxide, carbon monoxide, particulate

	Time Sensor Data				matter, and ozone. It also considered environmental variables such as wind speed and wind direction for the analysis
S2	A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction	Yang, Z. and Wang, J. (2017)	Pollutant data for two cities in China	China, specifically Xi'an and Jinan	PM10, PM2.5, and other major air pollutants
S3	Integrating Air Quality and Climate Change Strategies in Oman	Jehad Jabr Nasser Albusaidi, September 2019	National data for Oman, specifically data from 2010; includes emissions of various pollutants	Oman, with a focus on national strategies for air quality and climate change	Analysis of emission inventories, development of twelve emission scenarios including BAU and various mitigation stra
S4	Air Pollution Prediction using Machine Learning	Shreyas Simu, Varsha Turkar, Rohit Martires, Vranda Asolkar, Swizel Monteiro, Vaylon Fernandes, and Vassant Salgaoncar, 2020	The dataset consists of 1000 dummy data points generated using Gaussian distribution, covering parameters such as day, month, type of industry, size of the	the study involves air pollution data related to industrial emissions.	include day, month, type of industry, size of the industry, output efficiency of industry, and emission rate. Machine Learning algorithms like K-NN, SVR, RF, Multilinear Regression, and

			industry, and output efficiency of the industry		ANN were used for analysis.
S5	Public awareness, perceptions and attitudes on air pollution and its health effects in Muscat, Oman	Hilal K. Al-Shidi, Abdullah Khamis Ambusaidi & Hameed Sulaiman (2021)	A survey was conducted in Muscat, Oman between February and May of 2020.	The survey considered the disparity among gender, age, and education level of the respondents.	The study investigated four main aspects: sources of access to information, knowledge and risk perception about air pollution, and willingness to change and act for mitigation.
S6	A Machine Learning Model for Air Quality Prediction for Smart Cities	Usha Mahalingam, Kirthiga Elangovan, Himanshu Dobhal, Chocko Valliappa, Sindhu Shrestha, and Giriprasad Kedam, 2023.	Utilizes air quality data from Delhi, India	focused on urban environments, specifically the city of Delhi,	Employs Machine Learning Algorithms, specifically Neural Networks and Support Vector Machines, to predict the Air Quality Index (AQI), integrating data on various pollutants.

2.1Data Collection

The study dataset in this research was collected from Oman Environment Authority- Planning & Environmental Indicators Department between Jan 2023 and Dec 2023 from a station which located in Muscat area- Alkhwair in northers east side of Oman. The data set contains information about hourly emission limit values for various substances and environmental parameters. The details of all study variables with their units have been shown in Table 3.

Air quality index, representing the level of the contaminants inspected, not only measures these gasses (CO, NOX, SO2) but co also particulate matter. The sources of different pollutants and

their effects on human health are also varied. However, Traceability of pollution to specific activities, such as traffic or industrial emissions, is thus made feasible. The inclusion of methane (CH4) and non-methane hydrocarbons (NMHC) means that these can be indices of specific kinds of pollution sources or even tell us which type--for example agriculture or oil and gas-related activities. Measuring both PM10 and PM2.5 can enhance our knowledge on particulate matter, an important factor of air pollution that may affect health, especially the respiratory system.

Air quality data must be correlated with meteorological data[11] like temperature, humidity, wind speed, wind direction and atmospheric pressure. These factors affect the transport, change, and removal of pollutants from the atmosphere. For example, strong winds can disperse pollutants while Increased humidity levels may lead to the production of secondary pollutants. Solar radiation data can be particularly useful for photochemical studies, as sunlight drives the formation of ozone and other photochemical smog components.

Table 3. Description of Air pollution dataset attributes

Study Variables	Description	Unit of Measurement
CH4	Methane concentration	µg/m³
CO	Carbon Monoxide concentration	mg/m³
H2S	Hydrogen Sulfide concentration	µg/m³
NH3	Ammonia concentration	µg/m³
NMHC	Non-Methane Hydrocarbons concentration	µg/m³
NO	Nitric Oxide concentration	µg/m³
NOX	Nitrogen Oxides concentration	µg/m³
NO2	Nitrogen Dioxide concentration	µg/m³
O3	Ozone concentration	µg/m³
PA	Atmospheric Pressure	mbar
PM10	Particulate Matter (10 micrometres or less in diameter) concentration	µg/m³
PM25	Particulate Matter (2.5 micrometres or less in diameter) concentration	µg/m³
RAIN	Rainfall amount	mm
RH	Relative Humidity	%
SO2	Sulphur Dioxide concentration	µg/m³
SR	Solar Radiation	W/m²
TEMP	Temperature	°C
WD	Wind Direction	° (degree)

2.2 Data Acquisition

i. Air Quality Monitoring Stations: Stations with various sensors for different environmental pollutants and parameters provide real-time data of pollution and meteorological conditions.

ii. Meteorological Instruments: At our observatory sites there are instruments that provide weather information. This is the critical atmospheric data allowing us to see what specific environmental conditions result in elevated levels of smog.

The study collected the dataset from the Planning & Environmental Indicators Department at the Environmental Authority in the year 2023. Data was collected from a monitoring station located in Muscat area- Alkhwair, in the northeast side of Oman, between January 2023 and December 2023.

2.3 Data Preprocessing

Techniques such as mean imputation may have been used to fill in missing values. In mean imputation, missing values are replaced with the mean from all available data for that variable.

Z-score normalization is a way to make the data have a mean of zero and standard deviation one and so on, to facilitate finding outliers in terms of distance away from this mean value.

3. Methodology

This research experiment various machine learning models which is listed as follows

3.1 Regression Analyses

The purpose of regression analysis was to estimate pollution levels given environmental factors temperature, humidity, wind speed, and atmospheric pressure.

In a regression model, what this does is it builds a mathematical relationship between predictor variables (features) and dependent variables (pollution)[13]. For example, a linear regression model predicts pollution levels based on a linear equation that best fits the data.

Regardless of whether the model is linear or nonlinear, performance metrics including Mean Squared Error (MSE) and R-squared (R^2) are used to determine how well a model fits known data[24]. These metrics reflect the extent to which predicted pollution levels differ from actual observations.

3.2 Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) linear regression method is used to predict the relationship between one or more independent variables (predictors) and a dependent variable (response). The goal is to minimize the sum of squared differences between observed and predicted values. As long as the relationship is linear and additive, this method will be spot on[22]. Ordinary least squares regression is straightforward and easy to analyze, but it won't capture complex relationships in data.

3.3 Huber Regression Model

Huber's method of regression combines the merits of least squares and absolute deviation regression. It tries to minimize a combination of squared losses for small residuals and absolute losses for large residuals. When compared with ordinary least squares, Huber regression is not as sensitive to outliers, and it is thus suitable for datasets with outliers or non-constant variances[25]. In fact, it is both the reasonableness of absolute deviation regression and also the efficiency of least squares regression.

3.4 Multi-Layer Perceptron (MLP)

MLP is a kind of feedforward artificial neural network organized into layers. It has an input layer, one or more hidden layers, and an output layer. Each node in the hidden layers applies a nonlinear activation function to the weights and sum of its inputs. MLP is capable of learning complex nonlinear relationships in the data[26], but it may require tuning of hyper-parameters and is susceptible to overfitting.

3.5 Gradient Boosting

Gradient boosting is an ensemble learning technique that builds a strong predictive model by combining multiple weak models sequentially. It fits new models to the residuals (errors) of the previous ones, with each new model aiming to cut down on the errors made by its predecessors.In practice, gradient boosting is robust and can handle all types of data; but it may be computationally expensive and requires careful tuning of hyper-parameters.

3.6 Random Forests

Random forests are an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree in the forest is trained on a random subset of the training data and a random subset of the features. Random forests are not easy overfitting, and work well on high-dimensional data, providing estimates of feature importance[27].

Random forests are an ensemble learning method that produces many decision trees during training and outputs the mode of the classes (classification), or the average prediction (regression) of the individual trees. Breiman, L. (2002). Random forests and decision trees. Starting from scratch (p. 587). Shahid Beheshti medical teaching student. Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. R news, 2(3), 18-22.

3.7 Decision Tree

Decision trees are an individual method for supervised learning with nonparametric assumptions. It can be used for classification or regression tasks as well. In this way they divide up the feature space into regions, predicting class labels or values for each region by means of an assignment rule. Decision Trees are interpretable, easy to visualize, and can handle both numerical and categorical data. But they are susceptible to overfitting, especially when the trees get deep, and also may miss underlying complex patterns in the data.

3.8 K-Nearest Neighbors (KNN)

KNN is a lazy non-parametric classification and regression learning algorithm. It makes forecasts, then neighbors closest to the goal by use of majority voting (if classification) or averaging (if regression) the nearest k neighbors in the training data to decide on the class label or value for each new data point. KNN is quite simple and easy to implement--but it can also be very computationally expensive for large datasets. This algorithm is sensitive to the choice of distance metric and k value. Nearest neighbor pattern classification IEEE Transactions on Information Theory by Cover, T., & Hart, P. (1967).

3.9 Adaboost

The ensemble learner Adaboost combines many weak learners (usually decision trees) into a powerful classifier. The method trains weak models on subsets of the data one after another; while each subsequent model focuses on the misclassified instances of the previous models.

Among the misclassified instances Adaboost assigns higher weights to force the following models to place more emphasis on correcting these errors. It is more robust and not easy to overfat, compared with individual weak classifiers. by Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139. Friedman, J. H. (2001). On the asymptotic equivalence of weak and strong learning methods. Biometrika.

3.10 Support Vector Regression (SVR)

SVR is a support vector machine (SVM) and is used for regression. It determines the hyperplane that maximizes the margin between the predicted values and the actual values, subject to a specified tolerance (epsilon). SVR operates well in high-dimensional spaces, and is robust to outliers, but it may require tuning of hyper-parameters such as the kernel function and regularization parameter. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning by Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

3.11 Classification Analyses

In our study, the different types of classification analyses that were carried out. In classification, the goal of a supervised learning task is to predict the class or category to which a given input belongs based on its features. In our research, the following sections describe each of these classification analyses in more detail:

3.12 Evaluation Metrics

In order to gauge how well the classification models function, performance metrics were brought into play. Some typical performance metrics from your document are simple statistics like: Following evaluation metrics utilized to measure the performance of the model

3.12.1 Confusion matrix

The aim of the confusion matrix is to present the results of a classification model in a form that's easy to comprehend. A confusion matrix measures how a classification model is performing. This confusion matrix includes four important figures: true positives(TP), false positives(FP), true negatives(TN), and false negatives. False positives false negatives are counted to calculate performance measures such as accuracy, precision, recall, and F1-score[3].

In sum, the confusion matrix provides a detailed breakdown of a model's performance, including whether it can classify instances correctly or not, and so on.

A confusion matrix is a summary of the performance of the classification model and is often displayed as in Table 4. It includes four critical measures all together: True Positive, True Negative, False Positive, and False Negative.

Table 4. Confusion matrix

Classification	1 (High level of gas)	0 (Low level of gas)
1 (High level of gas)	TP	TN
0 (Low level of gas)	FN	FP

TP stands for the number of positive items that were correctly predicted, while TN is the number of negative items that were correctly predicted. FP is the number of incorrect positive

item predictions (Type I error), and FN is the number of wrong negative item predictions (Type II error).

The confusion matrix analysis is a detailed description of a model's performance, including its accuracy, precision, recall, and F1-score. Accuracy is defined as the proportion of instances that were correctly classified from all the instances (both high and low gas levels)[3]. Precision is the ratio of high gas levels which have been correctly predicted out of all that are predicted as high. Recall is the ratio of high gas levels correctly predicted among the actual number of high gas levels. The values are calculated as shown in Table 5.

Table 5. Key Performance Metrics for Binary Classification Models

Classification	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
AUC	Area under the curve from ROC

Receiver Operating Characteristic (ROC) Curves

ROC curves show the extent to which classification models can be distinguished between true positive and false positive. ROC curves delineate changes in sensitivity according to threshold settings. They plot the true positive rate (sensitivity) against the false positive rate (1 - specificity).

A model with a higher AUC (Area Under the Curve) value typically predicts better and does a better job distinguishing positive cases from negatives. Overall, the classification analyses are crucial to predicting pollution levels, and whether the pollution concentrations exceed threshold values. By trying out various classification algorithms and then using specific performance metrics to evaluate and rank these models, researchers can develop reliable models. These can help guide policy-making in air pollution control, and air quality management projects alike[7].

To illustrate how well an accurate classification model can distinguish a real positive from a false positive or negative rate, the graphing technique known as the receiver operating characteristic (ROC) curve was utilized[29]. These curves show how much the sensitivity ratio (i.e., true positive rate) and the specificity ratio (i.e., false negative rate) are varying between machines. The closer an ROC curve to the top left corner, the higher its predictive ability and the more it differentiates between positive and negative cases.

3.12.3 Area Under the Curve (AUC)

AUC is a quantification of the area under the ROC curve that plots the true positive rate (sensitivity) against false positive rates (1 - specificity) above various threshold settings. It quantifies the model's ability to distinguish high levels of gas from low, at the different thresholds[25].

4.Result & Discussion

Overall pollution pattern at each hour during a day is illustrated in Figure 8. The subplots show the dependency of CO with time, so that it gives one an idea of daily trends and peak position. Defining patterns is crucial when devising pollution control strategies to reduce the environmental and public health impacts. However, some further insights and interpretation about the observed daily and seasonal trends related to the pollutant may have been interesting. Perhaps, with the addition of some discussion on the possible reasons for these trends, such as varying traffic congestions, emergent industrial activities, or weather conditions, the understanding of the data can be enhanced. For instance, the possible reason for the increased CO at certain hours of the day or night might be related to increased vehicle emissions in peak traffic points of the day, especially during the winter season. Similarly, the seasonal variation in the pollutant might also be influenced by the climatic conditions of the region. The cold mountainous conduction, especially during night, might be trapping the pollutants to be monitored, instead of getting diffused into the sky or stratosphere. Such discussion would have added some extra value and produced a more complete analysis of the data.

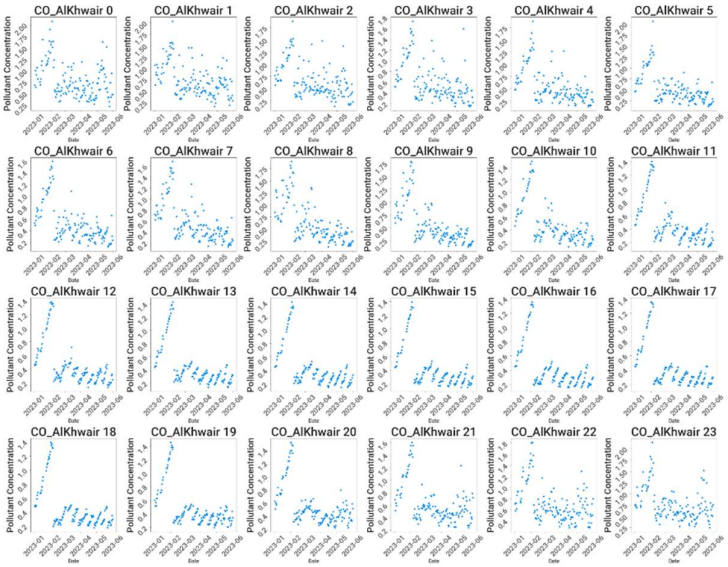

Figure 1. Overall pollution pattern at each hour during a day

The Table 6 below demonstrates the performance indicators for various regression models that were utilized to predict the total air pollution levels in Oman. These models were evaluated based on their mean squared error (MSE) and R2 score, which reveal the extent to which the independent variables can predict the dependent variable. Notably[30], the linear regression model stands out with an impressive R2 score and a remarkably low training MSE, indicating a strong fit to the training set. However, when tested on new data, it shows a slightly higher MSE and a significantly high R2 score, suggesting a possibility of overfitting.

“Low training MSE” in the context of provided information refers to an insignificant value of Mean Squared Error obtained while training the linear regression model. The training MSE is calculated as the average of the sum of squares of the difference between predicted and actual values of the dependent variable. Consequently, in terms of testing regression models, such an error rate reflects that obtained predictions closely correlate with the actual values in the given

model. The implication of this result is that the linear regression model has successfully fitted the training data, implying that it can predict outcomes in this dataset with high accuracy.

The Huber Regression approach is able to shine in generalization performance as it provides low MSE and high R^2 [17]values across our training and testing datasets, and is significantly more resilient to outliers and common data noise than ordinary least squares regression. The MLP (Multi-layer Perceptron) model also performs well with low MSE and high R^2 values across our datasets and seems to effectively capture complex nonlinear relationships in the underlying data. Gradient boosting is able to achieve strong predictive capabilities on our training data, with low MSE and high R^2 scores.

The results show that random forest model works very well. As seen based on the typical measures of performance on the training set, noticed the need to show cross-validated results indicating a very low mean squared error and perfect R^2. There is a noticeable increase in mean squared error and decrease in R^2 on the test data. Overfitting is a possibility. The decision tree model looks the best, with 0% MSE and a perfect R^2 score on the training set and a very high mean squared error and lower R^2 score on test data as would suggest it has severe overfitting[5].

Compared with other models, the K-Neighbors model has a high mean square error (MSE) and low R^2 scores on the training and testing datasets. It is suggested that this model may not be appropriate as an analytical tool for discerning the underlying patterns in the data. Not bad results for Ada Boost; on both training and testing sets it had medium R^2 scores, and a pretty high MSE. Ways to enhance its productivity are more features or optimization. On the training and testing sets SVR (Support Vector Regression) displayed a very high MSE and low R^2 values, indicating that it performs even more poorly than the K-Neighbors. Thus, the model may have difficulty in dealing with complex relationships within the data.

Table 6. Performance of proposed model for overall air pollution

Proposed Models	**Training MSE**	**Training R^2**	**Test MSE**	**Test R^2**
Linear Regression	0.000	1.000	0.001	1.000
Huber Regression	1.645	0.994	0.833	0.996
MLP	0.867	0.997	0.984	0.996
Gradient Boosting	1.533	0.994	4.009	0.982
Random Forest	1.334	0.995	5.246	0.977
Decision Tree	0.000	1.000	17.692	0.921
K-Neighbors	30.069	0.882	48.379	0.784
Ada Boost	87.338	0.656	88.854	0.603
SVR	225.047	0.113	195.660	0.125
Gaussian Process	0.000	1.000	242.500	0.084

The performance of several machine learning models for forecasting pollution indices, namely NMHC_AlKhwair, NO_AlKhwair, NOX_AlKhwair, NO2_AlKhwair, PA_AlKhwair, and PM25_AlKhwair, is shown by the Receiver Operating Characteristic (ROC) curves.

The ROC curves for NO_AlKhwair and NOX_AlKhwair show better performance than those for other contaminants. These curves are closer to the top left corner of the picture, embodying higher performance sensitivities and lower false-positive rates across various threshold settings. The models which best distinguish between positive and negative instances for NO_AlKhwair and NOX_AlKhwair outperformed the same method for other pollutant types. For NOX_AlKhwair and NO_AlKhwair, according to the ROC curves, they were most strongly correlated with pollution indicators of all kinds; this indicates that whatever models are being used have gained an especially good prediction accuracy at these pollutants. The concluding upshot is really, the model's features or predictors matter in more cases than we thought. We would also have to consider these specific contaminants-- their properties might suggest which features could enhance the accuracy of distinguishing among them and making predictions on data.

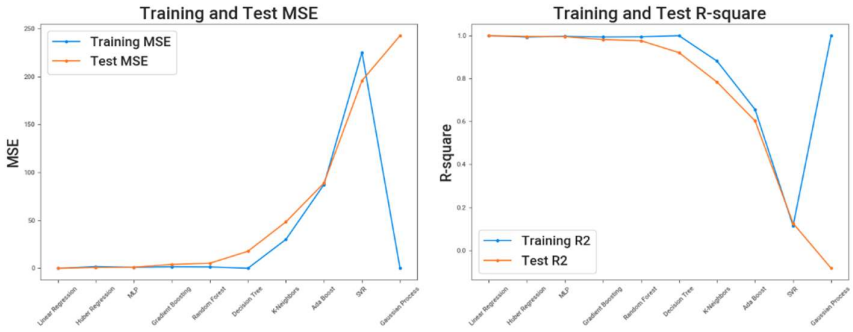

Figure 2. Performance of proposed models in terms of MSE and R^2

Hence, our fitted models offered valuable data and accurate predictions for NO_AlKhwair NOX_AlKhwair pollution levels and promoting public health.

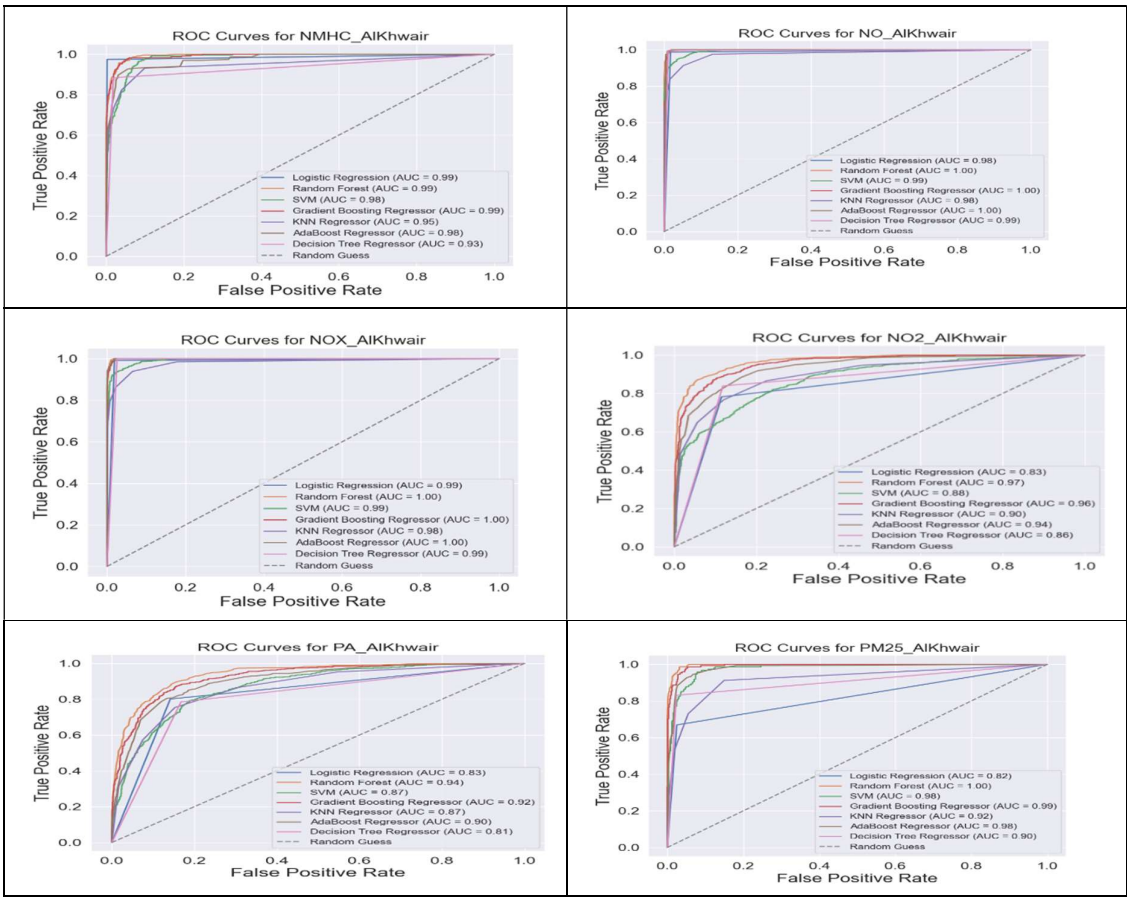

Figure 3.ROCs of all proposed models corresponding to pollution index (NMHC_AlKhwair, NO_AlKhwair, NOX_AlKhwair, NO2_AlKhwair, PA_AlKhwair, PM25_AlKhwair)

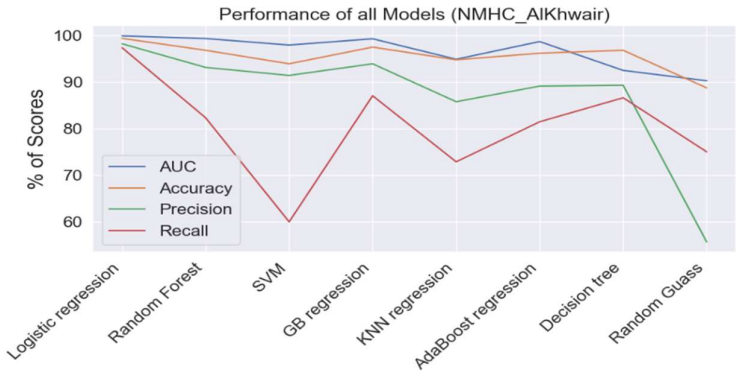

Figure 4. ML models performance in terms of AUC, Accuracy , Precision and Recall (NMHC_Alkhwair)

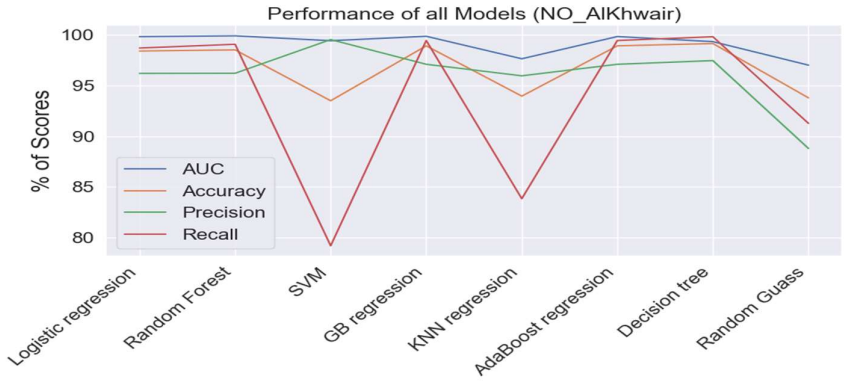

Figure 5. ML models performance in terms of AUC, Accuracy , Precision and Recall (NO_Alkhwair)

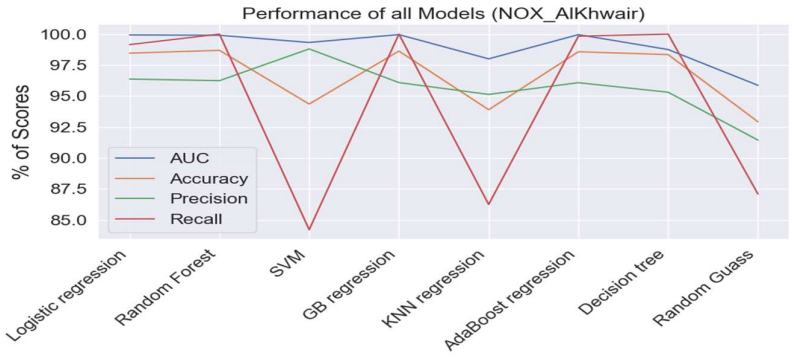

Figure 6. ML models performance in terms of AUC, Accuracy , Precision and Recall (NOX_Alkhwair)

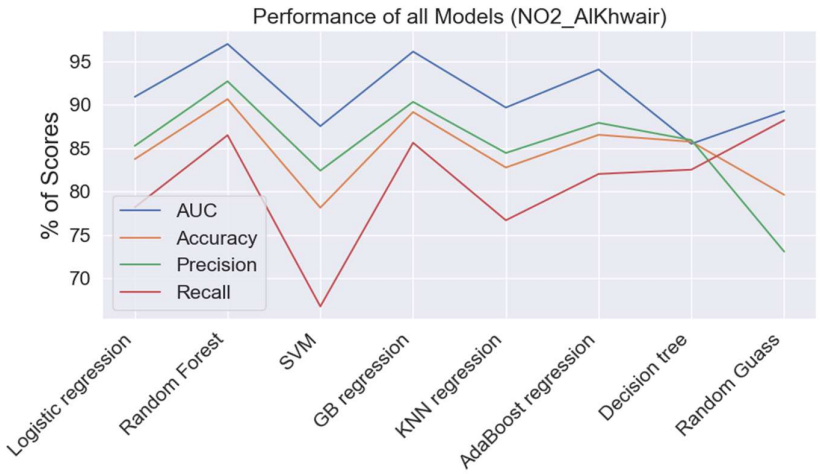

Figure 7. ML models performance in terms of AUC, Accuracy , Precision and Recall (NO2_Alkhwair)

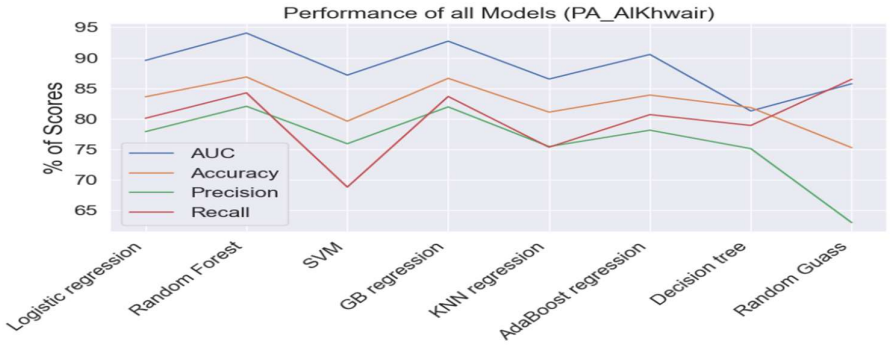

Figure 8. ML models performance in terms of AUC, Accuracy , Precision and Recall (PA_Alkhwair)

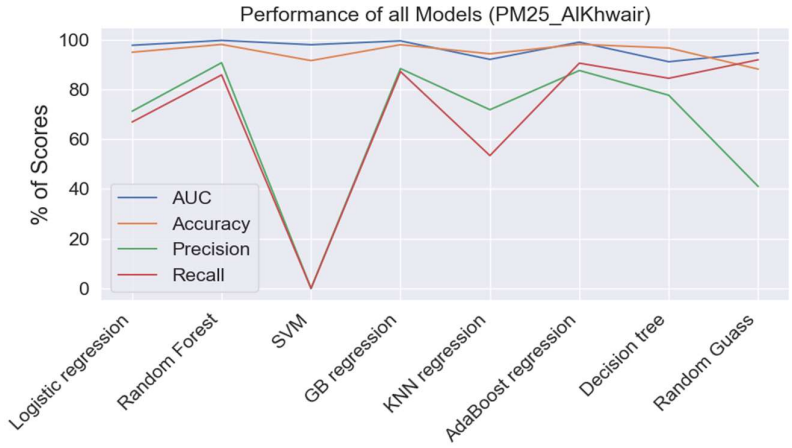

Figure 9. ML models performance in terms of AUC, Accuracy , Precision and Recall (PM25_Alkhwair)

Figure 4-9 shows the performance in terms of AUC, Accuracy , Precision and Recall for different features.

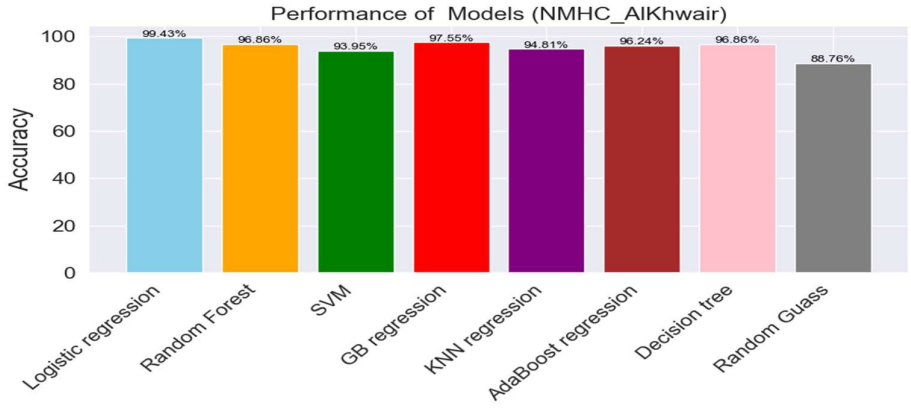

Figure 10. Accuracy score (NMHC_Alkhwair)for each of proposed ML models

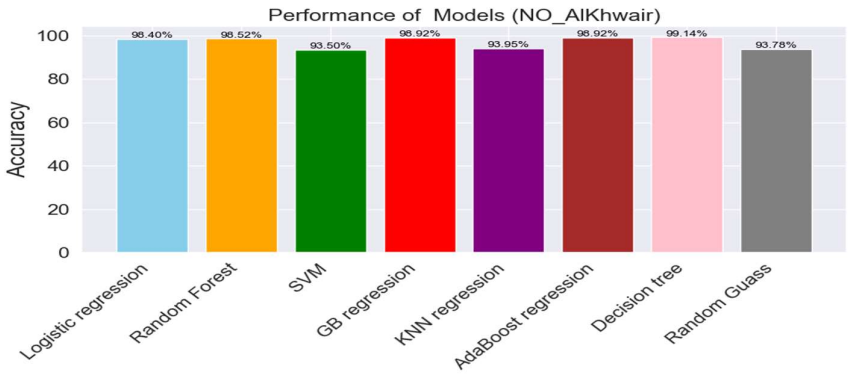

Figure 11. Accuracy score (NO_Alkhwair) for each of proposed ML models

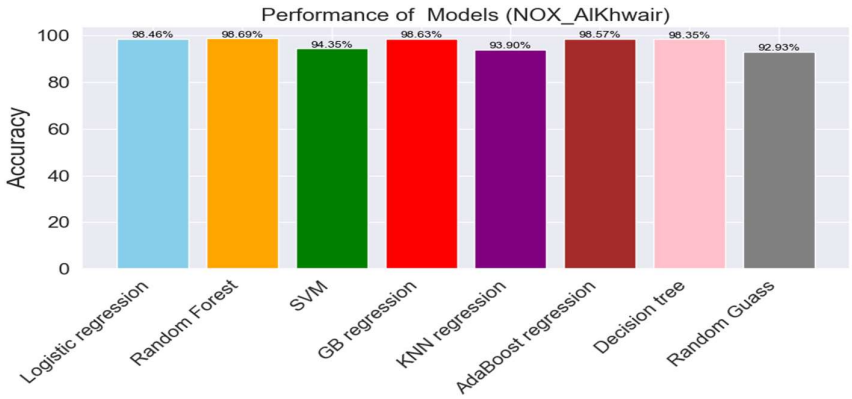

Figure 12. Accuracy score (NOX_Alkhwair) for each of proposed ML models

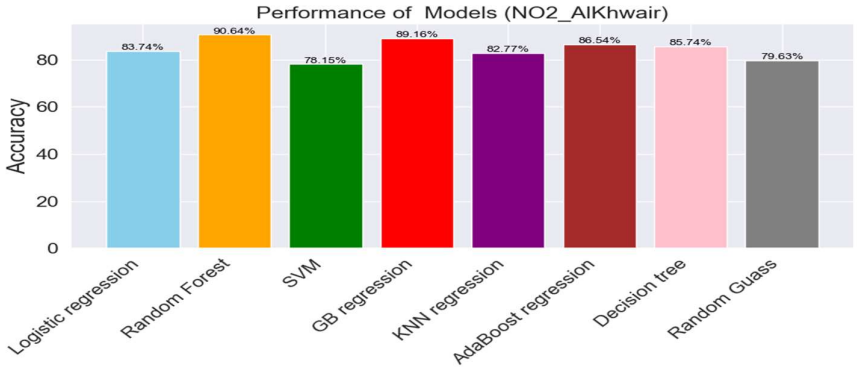

Figure 13. Accuracy score (NO2_Alkhwair) for each of proposed ML models

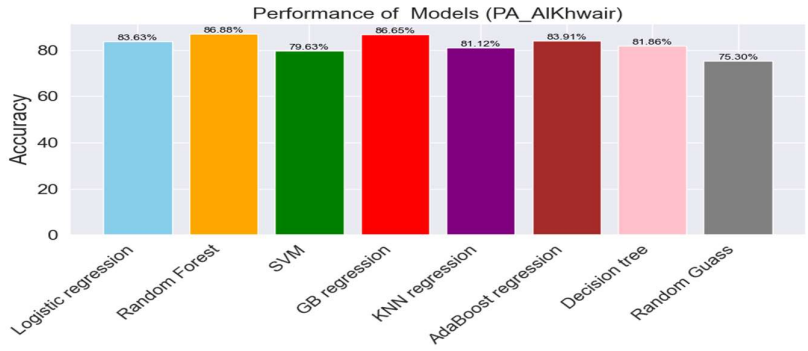

Figure 14. Accuracy score (PA_Alkhwair) for each of proposed ML models

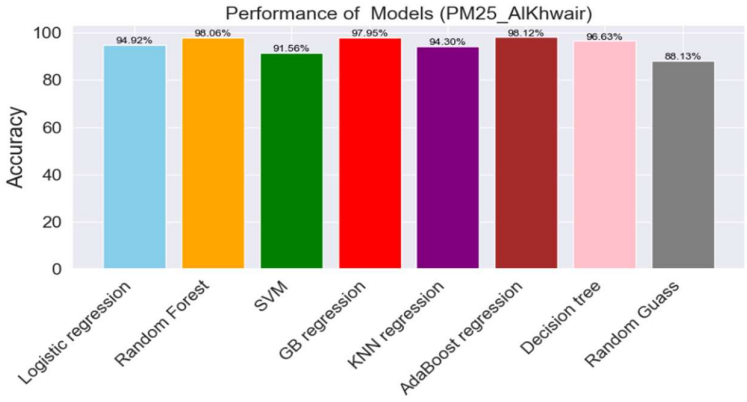

Figure 15. Accuracy score (PM25_Alkhwair) for each of proposed ML models

Figure 10 to 15 presents the Accuracy score of experimented models. The following figures 16-21 show the performance of machine learning models in terms of Accuracy and AUC for each of the pollutants individually. A detailed explanation is provided for each figure so that the reader is able to understand the performance of machine learning models for different pollutants in forecasting air pollution indices.

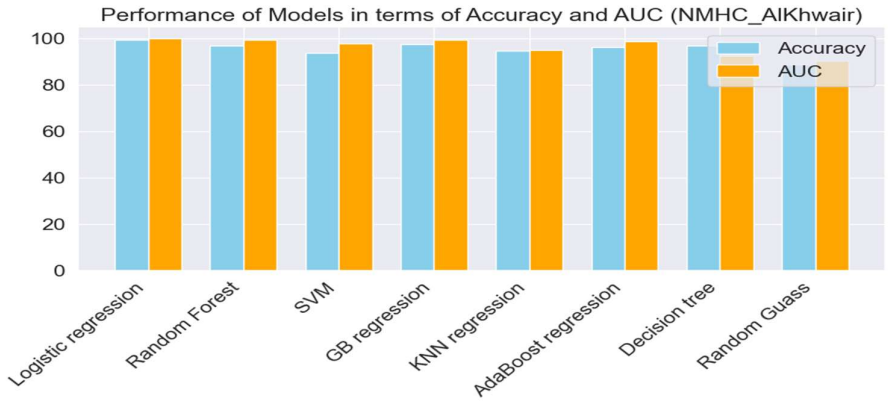

Figure 16. Analysis of ML model's performance in terms of Accuracy and AUC (NMHC_Alkhwair)

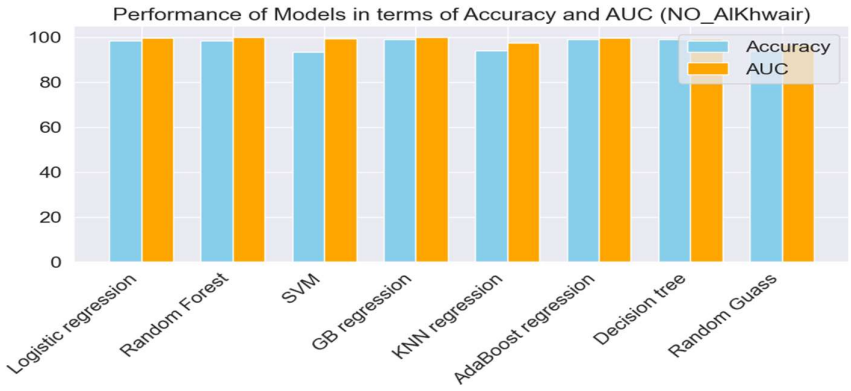

Figure 17. Analysis of ML model's performance in terms of Accuracy and AUC (NO_Alkhwair)

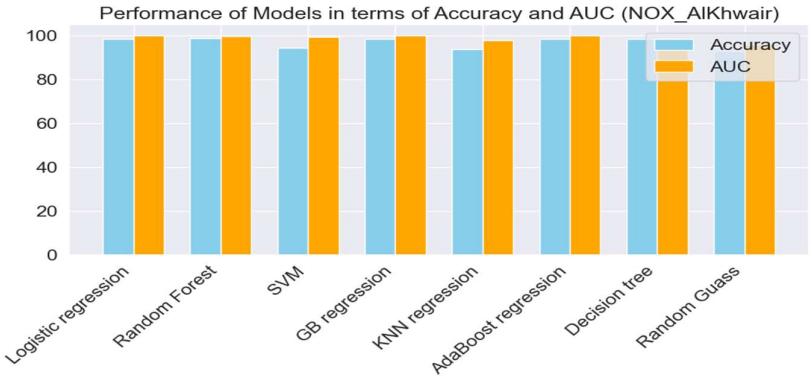

Figure 18. Analysis of ML model's performance in terms of Accuracy and AUC (NOX_Alkhwair)

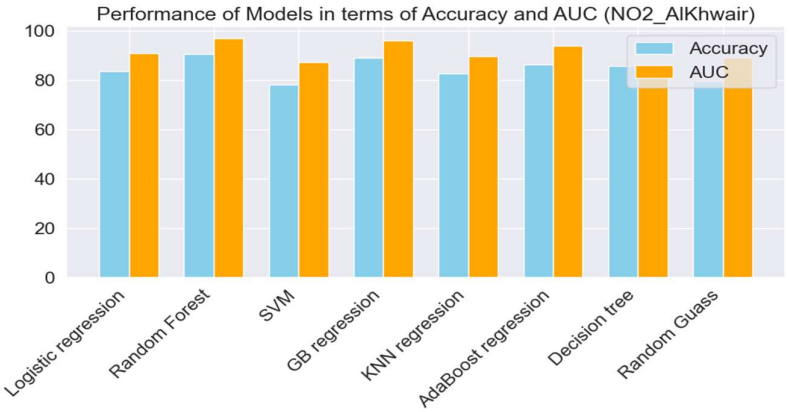

Figure 19. Analysis of ML model's performance in terms of Accuracy and AUC (NO2_Alkhwair)

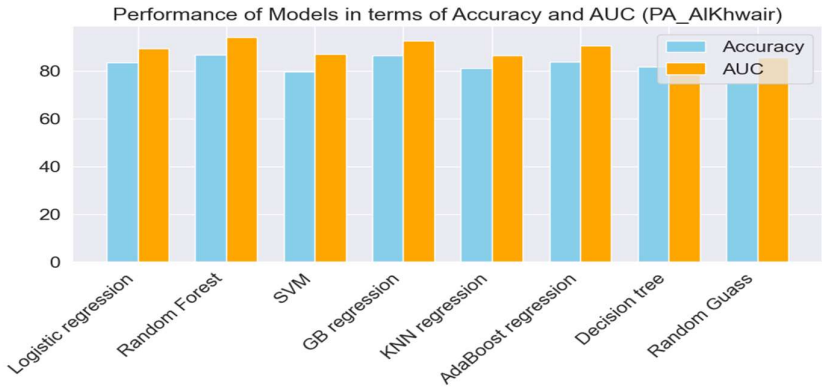

Figure 20. Analysis of ML model's performance in terms of Accuracy and AUC (PA_Alkhwair)

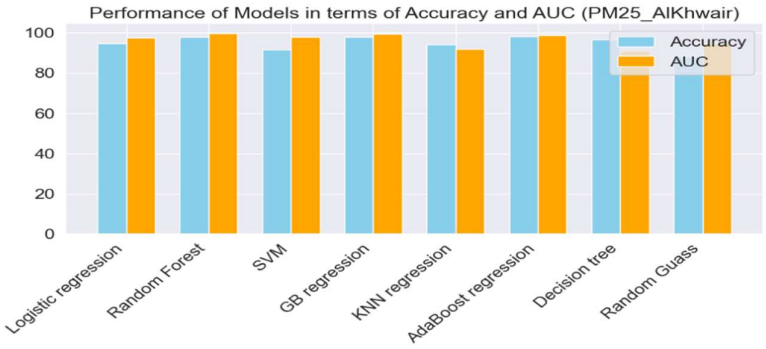

Figure 21. Analysis of ML model's performance in terms of Accuracy and AUC (PM25_Alkhwair)

Conclusion

The research sought to demonstrate the application of machine learning in predicting air pollution levels in Oman by implementing regression and classification models. We set out to investigate how effectively the algorithms could predict pollutant concentrations; and binary pollution levels, which can provide important input to environmental management and public health initiatives. By assessing the performance of regression models like Linear Regression, Huber Regression and MLP, and classification models such as Logistic Regression and Random Forest, we were able to uncover trends revealing the most effective means of forecasting air pollution indices. In addition, we demonstrated nuances in model performance across different pollutants, highlighting the potential benefits of customized approaches for specific contaminants. Overall, this study indicates that machine learning has the potential to significantly enhance the prediction of air quality and suggests several directions for future research and model enhancements that could lead to improved ability to address the challenges of air pollution in Oman.

References

[1] A. Pandey *et al.*, "Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019," *Lancet Planet Health*, vol. 5, no. 1, pp. e25–e38, Jan. 2021, doi: 10.1016/S2542-5196(20)30298-9.

[2] Z. Yang and J. Wang, "A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction," *Environ Res*, vol. 158, pp. 105–117, 2017, doi: 10.1016/j.envres.2017.06.002.

[3] T. Sable, S. Bodkhe, H. Bhamare, P. Bhangare, and U. Student, "Air Quality Monitoring and Prediction System," *Int J Innov Res Sci Eng Technol*, 2020, [Online]. Available: www.ijirset.com

[4] J. Bachmann, "Air pollution forecasts and results-oriented tracking," *Air Qual Atmos Health*, vol. 1, no. 4, pp. 203–207, 2008, doi: 10.1007/s11869-008-0025-8.

[5] P. Bhalgat, "Air Quality Prediction using Machine Learning Algorithms," 2019. [Online]. Available: www.ijcat.com

[6] "A consumer application that demonstrates how the cost of living has changed in the UK since 2008, based on lifestyle choices".

[7] M. Said, K. ben Abdellafou, O. Taouali, and M. F. Harkat, "A new monitoring scheme of an air quality network based on the kernel method," *International Journal of Advanced Manufacturing Technology*, vol. 103, no. 1–4, pp. 153–163, Jul. 2019, doi: 10.1007/s00170-019-03520-9.

[8] "A Machine Learning Approach to Improving Staff Shift Scheduling in Home Healthcare_Redacted".

[9] R. Haghbakhsh, H. Adib, P. Keshavarz, M. Koolivand, and S. Keshtkari, "Development of an artificial neural network model for the prediction of hydrocarbon density at high-pressure, high-temperature conditions," *Thermochim Acta*, vol. 551, pp. 124–130, Jan. 2013, doi: 10.1016/j.tca.2012.10.022.

[10] L. T. Ator *et al.*, "Determination of Air Pollutant Concentrations in Plant Species in Relation to Pollution Sources," *Open Journal of Air Pollution*, vol. 10, no. 03, pp. 53–62, 2021, doi: 10.4236/ojap.2021.103004.

[11] H. A. Al-Jamimi, S. Al-Azani, and T. A. Saleh, "Supervised machine learning techniques in the desulfurization of oil products for environmental protection: A review," *Process Safety and Environmental Protection*, vol. 120. Institution of Chemical Engineers, pp. 57–71, Nov. 01, 2018. doi: 10.1016/j.psep.2018.08.021.

[12] W. Sun and M. Liu, "Prediction and analysis of the three major industries and residential consumption CO2 emissions based on least squares support vector machine in China," *J Clean Prod*, vol. 122, pp. 144–153, May 2016, doi: 10.1016/j.jclepro.2016.02.053.

[13] D. Patel and J. I. Nirmal Kumar, "An Evaluation of Air Pollution Tolerance Index and Anticipated Performance Index of Some Tree Species Considered for Green Belt Development: A Case Study of Nandesari Industrial Area, Vadodara, Gujarat, India," *Open Journal of Air Pollution*, vol. 07, no. 01, pp. 1–13, 2018, doi: 10.4236/ojap.2018.71001.

[14] E. Cancila, I. Sabbadini, M. Ottolenghi, M. Deserti, and S. Tessitore, "Citizens and Air Quality: The Results of the First Survey Carried Out in the Po River Basin (Northern Italy)," *Open Journal of Air Pollution*, vol. 08, no. 03, pp. 69–79, 2019, doi: 10.4236/ojap.2019.83003.

[15] Y. Rybarczyk and R. Zalakeviciute, "Assessing the COVID-19 Impact on Air Quality: A Machine Learning Approach," *Geophys Res Lett*, vol. 48, no. 4, Feb. 2021, doi: 10.1029/2020GL091202.

[16] "Air pollution." Accessed: Apr. 07, 2024. [Online]. Available: https://www.who.int/health-topics/air-pollution#tab=tab_1

[17] S. G. Gunjal and S. Kamalapurkar, "Air Pollution Prediction Using Machine Learning Supervised Learning Approach", [Online]. Available: www.ijstr.org

[18] Y. C. Liang, Y. Maimury, A. H. L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," *Applied Sciences (Switzerland)*, vol. 10, no. 24, pp. 1–17, Dec. 2020, doi: 10.3390/app10249151.

[19] IEEE Singapore Section, IEEE Region 10, and Institute of Electrical and Electronics Engineers, *Proceedings of the 2016 IEEE Region 10 Conference (TENCON) : November 22-25, 2016, Marina Bay Sands, Singapore.*

[20] S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani, and N. Niveshitha, "Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques," *IEEE Access*, vol. 11, pp. 115140–115149, 2023, doi: 10.1109/ACCESS.2023.3323447.

[21] IEEE Singapore Section, IEEE Region 10, and Institute of Electrical and Electronics Engineers, *Proceedings of the 2016 IEEE Region 10 Conference (TENCON) : November 22-25, 2016, Marina Bay Sands, Singapore.*

[22] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: An overview," *International Journal of Environmental Research and Public Health*, vol. 15, no. 4. MDPI AG, Apr. 17, 2018. doi: 10.3390/ijerph15040780.

[23] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, May 2023, doi: 10.1007/s13762-022-04241-5.

[24] T. A. Saleh, K. O. Sulaiman, S. A. AL-Hammadi, H. Dafalla, and G. I. Danmaliki, "Adsorptive desulfurization of thiophene, benzothiophene and dibenzothiophene over activated carbon manganese oxide nanocomposite: with column system evaluation," *J Clean Prod*, vol. 154, pp. 401–412, Jun. 2017, doi: 10.1016/j.jclepro.2017.03.169.

[25] K. Dubey, S. Verma, S. Santra, and M. Kumar, "Identification of Critical Locations for Improvement of Air Quality Developing a Prioritized Clean Air Assessment Tool (PCAT)," *Urban Science*, vol. 7, no. 3, Sep. 2023, doi: 10.3390/urbansci7030075.

[26] Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 2057–2066. doi: 10.1016/j.procs.2020.04.221.

[27] "Air Quality Prediction in East Bay Area, CA Building a Machine Learning Model for Air Quality Predictions." [Online]. Available: https://daymet.ornl.gov/

[28] S. Suganya and T. Meyyappan, "Prediction of the level of air pollution using adaptive neuro-fuzzy inference system," *Multimed Tools Appl*, vol. 82, no. 24, pp. 37131–37150, Oct. 2023, doi: 10.1007/s11042-023-15046-0.

[29] Institute of Electrical and Electronics Engineers, IEEE ITSS, and T. ICT4ALL (Conference) (2015 : Ḥammāmāt, *10th IEEE Int. Conf. on Service Operations and Logistics, and Informatics : SOLI 2015 : November 15-17, 2015, Yasmine Hammamet, Tunisia.*

[30] V. Thamilarasi, P. K. Naik, I. Sharma, V. Porkodi, M. Sivaram and M. Lawanyashri, "Quantum Computing - Navigating the Frontier with Shor's Algorithm and Quantum Cryptography," 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, Pune, India, 2024, pp. 1-5, doi: 10.1109/TQCEBT59414.2024.10545283.

[31] T. Madan, S. Sagar, and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms-A Review," in *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 140–145. doi: 10.1109/ICACCCN51052.2020.9362912.

[32] K. Mukhopadhyay, R. Ramasamy, B. Mukhopadhyay, S. Ghosh, S. Sambandam, and K. Balakrishnan, "Use of Ventilation-Index in the Development of Exposure Model for Indoor Air Pollution—A Review," *Open Journal of Air Pollution*, vol. 03, no. 02, pp. 33–41, 2014, doi: 10.4236/ojap.2014.32004.

[33] Z. Zhang, C. Johansson, M. Engardt, M. Stafoggia, and X. Ma, "Improving 3-day deterministic air pollution forecasts using machine learning algorithms," *Atmos Chem Phys*, vol. 24, no. 2, pp. 807–851, Jan. 2024, doi: 10.5194/acp-24-807-2024.

[34] "Analysing the environmental impacts of recipe box recipes with recommendations for improving the sustainability of customer choices (1)".

[34] S. Finardi, R. De Maria, A. D'Allura, C. Cascone, G. Calori, and F. Lollobrigida, "A deterministic air quality forecasting system for Torino urban area, Italy," *Environmental Modelling and Software*, vol. 23, no. 3, pp. 344–355, Mar. 2008, doi: 10.1016/j.envsoft.2007.04.001.