Lahore's Air Quality Revealed: A Data-Driven Analysis Using Supervised Learning

Aleena Shafqat Butt[1], Mahnoor Irfan[2] and Nadia Mushtaq[3]

https://doi.org/10.62345/jads.2025.14.2.34

Abstract

Ensuring safe and clean air quality for breathing is crucial as it is getting effected by several factors. This study compares data taken at midnight and afternoon to find the temporal differences and possible contributing factors to examine the Air quality in Lahore. Air quality indices (AQI) and environmental variables are modelled using machine learning (ML) models (i.e. Linear regression, decision tree, random forest, SVM, KNN, Gradient boosting, Lasso regression, Ridge regression.) to study pattern variations in the quality of the air between the two periods, which is indicative of the impact of atmospheric dynamics. Classification is being done on the data using machine learning models which have been evaluated based on precision, accuracy, recall etc. Comparison of both afternoon and midnight data have been studied and by the result we have made the conclusion that at midnight AQI is worse than the morning AQI. Decision-makers can use these findings to improve public health and protect the environment. In both periods, $PM_{2.5}$ and PM_{10} turn out to be the most favorable predictors of AQI for both periods (Afternoon: Nighttime: PM10 B = 45.30, p <.05; Midnight: PM10 B = –86.2, p <.05, whereas NO_2 has a negative association in the afternoon model (B = –29.48, p <.05). There is a significant difference between mean AQI at midnight (M=223.9) and afternoon (M=194.5), which means that nocturnal air quality is poorer. In regression techniques, Lasso is the best performing method during afternoon (minimum MSE= 1567.2, maximum R^2 = 0.24 for classification tasks). simple linear regression works best at midnight, decision tree and random forest classifiers are better than SVM and KNN in attaining a perfect accuracy (100%) or classification of AQI levels ("Poor", "Unhealthy", "Severe") at 00:00 hours and 12:00 hours. Based on these results, it can be seen how Lahore air pollution changes throughout the day and how PM becomes of the highest significance for the overall AQI.

Keywords: Lahore Air Quality, Supervised Learning, Pollutant, SVM, KNN, Random Forest.

Introduction

The AQI serves as a vital role for communicating the air quality of a specific location, providing a clear understanding of measures of pollution levels. In cities like Lahore levels of air pollution have increased over recent years due to the factors i.e. Industrialization, rapid growth in population, vehicles etc.

[1]*Department of Statistics, Forman Christian College (A Chartered University), Lahore, Pakistan.*
[2]*Department of Statistics, Forman Christian College (A Chartered University), Lahore, Pakistan.*
[3]*Department of Statistics, Forman Christian College (A Chartered University), Lahore, Pakistan.*
Corresponding Author Email: nadiamushtaq@fccollege.edu.pk

Compliance with ethical standards: There are no conflicts of interest (financial or non-financial). This study did not receive any funding.

The air quality of Lahore, the largest metropolitan city, regularly surpasses the threshold set by the international community and that of WHO, ranking this city as one of the dirtiest cities globally. Some health risks, like respiratory and cardiac diseases, prevail, and they also impact the ecological and economic health of the region.

Understanding the temporal patterns of air quality is important to identify episodic pollution events, seasonal trends, and diurnal fluctuations. These trends inform how human activity, industrial processes, and weather patterns impact air pollution. For example, smog has become a persistent environmental problem in Lahore during the winter months, exacerbated by temperature inversions and crop leftovers being burned. Concentrations of pollutants also show observable diel cycles caused by temporal variation in industrial emissions and traffic intensity.

The temporal dynamic of air quality in Lahore is examined based on the critical pollutants: O_3, SO_2, Co, NO_2, and PM2.5 and PM10. This study aims to spot patterns by looking into the temporal fluctuations of these important parameters that contribute to them to be able to enlighten planners and policymakers to construct efficient plans for air quality control. To cut down pollution levels and ensure a better, sustainable urban environment for the people of Lahore, these trends must be understood.

Ozone (O_3) is created by photochemical interactions between sunlight, nitrogen oxides (NO_x), and volatile organic compounds (VOCs), primarily from industrial processes and automobile emissions. It can worsen asthma, impair lung function, and cause respiratory problems.

The main sources of Sulphur dioxide (SO_2), which causes respiratory issues, eye irritation, and an elevated risk of cardiovascular diseases, are the combustion of fossil fuels, industrial operations, and power plants.

The incomplete combustion of automobiles and burning fuels produces carbon monoxide (CO), which lowers oxygen delivery in the body and can cause headaches, lightheadedness, and even deadly poisoning at high amounts.

Nitrogen Dioxide (NO_2), which is created by industrial processes and vehicle emissions, affects lung function, increases vulnerability to infections, and causes respiratory system inflammation.

Looping back to the particulate matter (PM2.5 and PM10) – dust, building sites, car exhaust, industrial pollution, which may enter the bloodstream and lungs to a great depth and induce serious respiratory condition, heart issues, and even early death. Together, these pollutants form a serious public health crisis in Lahore, further exacerbating respiratory and cardiovascular problems in the locals.

A worldwide health risk, ambient particulate matter ($PM_{2.5}$) with 4.2 million premature deaths per year represented is believed to be responsible for this burden with the South Asia being the most affected by this burden (WHO, 2021).

One of the leading countries in terms of the highest annual mean $PM_{2.5}$ levels in 2024 were Pakistan. The annual mean in Lahore exceeded 100 µg/m³, which is more than ten times more than the WHO recommendation of 10 µg/m³.

In Punjab's urban areas, as much as a 60%increase in $PM_{2.5}$ levels has been observed during winter months i.e. November- December due to temperature inversions and burning of crop residues. From the preliminary data gathered between October and December 2024, the midnight air quality index in Lahore is usually 15% larger than the afternoon air quality index (M = 223.9, SD = 48.4; M = 194.5, SD = 49.9), which corresponds to a lack of mixing.

The last three years have witnessed consistently imminent dangerous air quality measurements in Lahore, during the winter. In November 2024, the $PM_{2.5}$ concentrations in the city were recorded at 125.2 times the annual air quality guideline value set by the WHO, leading to an all-time high AQI of 1,165 at the early hours of the day.

In a similar vein, Lahore was the most polluted city in the world in November 2023 when the AQI of the city rose to a harmful level of 470. Vehicle emission, industrial activities, burning of leftovers from crops, and adverse weather conditions that concentrate the pollutants near the surface are the primary causes of such high levels of pollution.

Lahore's comprehensive and persistent air pollution highlights the need to understand temporal pollution flows and activation of effective mitigation response. This study aims at exploring the diurnal changes in Lahore's air quality to inform focused policy measures and amplify public health outcomes, as well as to focus on the differences between midnight and afternoon AQI readings.

Air pollution in smart cities has been adversely affecting the bodily health and standard of living of people. Therefore, reliable air quality forecasting is an essential part of a successful action plan for air pollution reduction and construction of more sustainable healthy ecosystems (Al-Eidi et al., 2023). Levels of air pollution and their impact on public health are monitored with the help of an Air Quality Index (AQI). Because of the increasing number of automobiles and greater industrialization that have made air pollution problems worse, more accurate AQI projections are needed (Ct & Amutha, 2024). Industrialization, population expansion, traffic and weather patterns. Also, the use of crude oil derivatives and other mechanical elements all contribute to air pollution, which has a negative impact on living things.

These include pollutants like PM2.5, PM10, NO2, SO2, O3, CO that mainly affect the air quality [(Suroshe et al., n.d.),(Natarajan et al., 2024)] and DELYS and deaths by air quality examined by (Amer et al., 2024). Urbanization and industrialization make it tough to rapidly identify and manage air pollution sources. Machine learning techniques have been successfully used in tackling air pollution issues by effective tracking of emission sources (Choi et al., 2024). The objective of the AQI is to create public awareness of the harmful health impacts of local air pollution. There are several approaches in developing a mathematical method to calculate the air quality index. Among the most interesting methods for predicting and analyzing AQI are data mining techniques (Gupta et al., 2023). It can be approached innovatively by integrating machine learning techniques with secondary data modeling (Liu et al., 2024). Various methods are proposed in the literature and employed to develop models to analyze how the air quality index varies with time. This also involves the development of a multiple linear regression model from the available data on the air quality index (Loganathan et al., 2022). In Lahore, Pakistan, smog has become a major environmental problem that is endangering agricultural productivity and aggravating health issues which emphasized how air conditions, burning crop residue, vehicle pollution, and industrial pollutants interact (Razzaq et al., 2024).

In many cities, especially other developing countries like India, air pollution has become a serious concern. Ambient air quality needs to be maintained by frequent air pollution forecasting and monitoring. Unlike traditional methods, machine learning has emerged as a feasible approach to predict the Air Quality Index (AQI) [(Ravindiran et al., 2023),(Nataraj & Goldsztein, n.d.)]. The air quality of different regions like Macao must be predicted precisely. Machine learning techniques including random forest (RF), gradient boosting (GB), support vector regression (SVR), and multiple linear regression (MLR) were applied to predict the concentration of particulate matter (PM10 and PM2.5) in those certain regions (Lei et al., 2022). The machine learning techniques are also used in predicting air quality and can be classified by using different machine learning models. Thus, predictive metrics of air quality in severely affected locations are

therefore critical towards enhancing quality of life, with the predominant focus being on PM10 values of the air quality (Mampitiya et al., 2023). Since air pollution is so subtle, it has been called a silent killer. Its harmful effects are further emphasized by the indirect effect it has on human health. Early air quality detection may save millions of lives worldwide (Rahman et al., 2024) .
The purpose of this study is to identify the day with the worst air quality by examining the changes in the AQI between midnight and afternoon. We aim to check air pollution monitoring and find patterns by using supervised learning models for regression and classification. The results will help raise public knowledge of Lahore's air pollution patterns and support data-driven policymaking.

Methodology

In this we have considered the methods related to Statistical algorithms and machine learning models for the analysis of air quality in Lahore when it was at its peak. We have used different machine learning models such as Linear regression, Ridge regression, Lasso regression, Decision tree, Random Forest, Support-vector machine, KNN and gradient boosting. Each of these machine learning techniques has specific advantages and is applied to both regression and classification applications. Continuous values are predicted using linear, ridge, and Lasso regression; Lasso selects features by reducing some of the coefficients to zero, while Ridge manages multicollinearity. Random forests improve accuracy by merging several trees to lessen overfitting, whereas decision trees provide interpretable models by dividing data according to feature values. By identifying the best decision boundaries, Support Vector Machines (SVMs) are effective for classification, particularly in high-dimensional domains. A straightforward, non-parametric technique that depends on the proximity of data points is K-Nearest Neighbor (KNN). Gradient boosting uses many weak learners to gradually correct errors, increasing prediction accuracy. These techniques are generally applicable to real-world issues since they accommodate different data complexities.

Linear Regression

A simple but effective algorithm for supervised learning used to predict a continuous dependent variable given one or more independent variables is Linear regression It takes the assumption of linear relation of the input variables (features) to the target variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{1}$$

Here,
y= dependent variable.
x= independent variable.
β_0= intercept.
β_1= slope of the line (coefficient).
ε= error term.

Ridge regression

It is quite like least squares except it estimates the coefficients by minimizing a slightly different quantity. The riding regression coefficient, in particular, the values that minimize are the calculations that are made.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 = \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} {\beta_j}^2 \tag{2}$$

Here,

$\lambda \geq 0$is a tuning parameter, to be determined separately.
$\lambda \sum_{j=1}^{p} \beta_j$is called shrinking penalty, it is small when $\beta_1, \dots, \beta_p$are close to zero and so it has the effect of shrinking the estimates $\beta_j$towards zero.

Lasso regression
In ridge regression the penalty $\sum \beta_j{}^2$ will shrank the coefficients towards zero but it will not set any of them equal to zero (unless $\lambda = infinity$) This may not be a problem for prediction but it can become a challenge for model interpretation especially when number of variables p is quite large. So, to handle this problem, Lasso regression is used.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j\, x_{ij}\right)^2 = \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3)$$

Decision tree
Decision Tree is a supervised learning algorithm used in both classification, regression, and it is one of the non-parametric algorithms. It uses conditions for value of features to branch information into a tree structure.
It includes two categories:

Gini impurity

$$G = 1 - \sum_{i=1}^{n} p_i^2 \quad (4)$$

Here, $p_i = \frac{Number\ of\ instances\ in\ class\ i}{total\ number\ of\ instances\ at\ the\ node}$

Entropy

$$H = -\sum_{i=1}^{n} p_i log_2(p_i) \quad (5)$$

Here,
k= number of classes.
p_i=propotion of instances in class i.

Random forest
In random forest we combine multiple decision trees to increase accuracy and decrease overfitting.
For classification:

$$\hat{y} = mode(y_1, y_2, \dots, y_k) \quad (6)$$

For regression:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \quad (7)$$

Support vector machine (SVM)
It is a potent learning model for both regression and classification. The goal is to determine the best hyperplane for either classifying data points into distinct classes or predicting a continuous output. In classification, SVM is resilient against overfitting in high-dimensional spaces since it maximizes the margin between the classes.

$$L(w, b, \epsilon) = \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} (\xi i + \xi i *) \quad (8)$$

Here,
$\parallel w \parallel^2$= is the regularization term it controls the complexity of the program.

ξi and ξi^* represent the deviation from the margins.
C is the regularization parameter that determines the trade-off between margin size and slack penalties.

KNN

KNN regression is a simple instance-based approach that averages the values of a data point's k-nearest neighbors to forecast the target value for that data point. A distance metric such as Euclidean distance is used to calculate the proximity of neighbors.
Prediction for a query point

$$\hat{y} = \frac{\sum_{i \in Nk} wiyi}{\sum_{i \in Nk} wi} \tag{9}$$

Here
N_k=set of k nearest neighbor of x.
wi= weighted assigned to the neighbor i.

Gradient boosting

A machine learning method for classification and regression issues is called gradient boosting. It creates an ensemble of weak learners, usually decision trees, to create a predictive model. The model is trained relatively by using gradient descent to minimize a given loss function, with each new model seeking to fix the mistakes caused by the ones before it.
First, we will initialize the model:

$$F_0(x) = argmin_c \sum_{i=1}^{n} L(y_i, c) \tag{10}$$

Here, L is a loss function, c is a constant that minimizes the loss function across the data set, mean we have to find the predicted values for which the loss function is minimum.
For each iteration m= 1,2, 3, …, M:

$$r_i^{(m)} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1(x)}} \tag{11}$$

Here,
$y_i$is the observed value.
r is residual.
m is the no. of iterations we are performing.

Data source and Sample collection

The data for this study on Lahore's air quality was collected from https://www.aqi.in/dashboard/pakistan/punjab/lahore, capturing main air quality monitors such as PM2.5, PM10, CO, SO2, NO2, and O3. Measurements were recorded at two distinct time periods—midnight and afternoon—to examine temporal variations in pollution levels. The data set includes daily recorded observation from 20-10-2024 to 20-12-2024. The data was recorded two times Midnight and Afternoon.
The Air Quality Index (AQI) measurements in Lahore experienced major ups and downs during October 20 through December 20, 2024, showing big peaks occurring during daytime and at midnight. Lahore experienced hazardous air conditioning when its AQI reached 708 at 11 PM on October 29. The lowest Air Quality Index reading for that entire day remained at level 246 during the 4 to 5 PM hour before ending as extremely dangerous. Lahore achieved dangerously high AQI scores that ranked it among the world's most polluted cities throughout November when nighttime conditions worsened the situation. Vehicle exhaust and industrial operations together with freezing

atmospheric conditions that restrained air pollution to lower elevations caused increased AQI readings in the late afternoon and throughout nighttime.

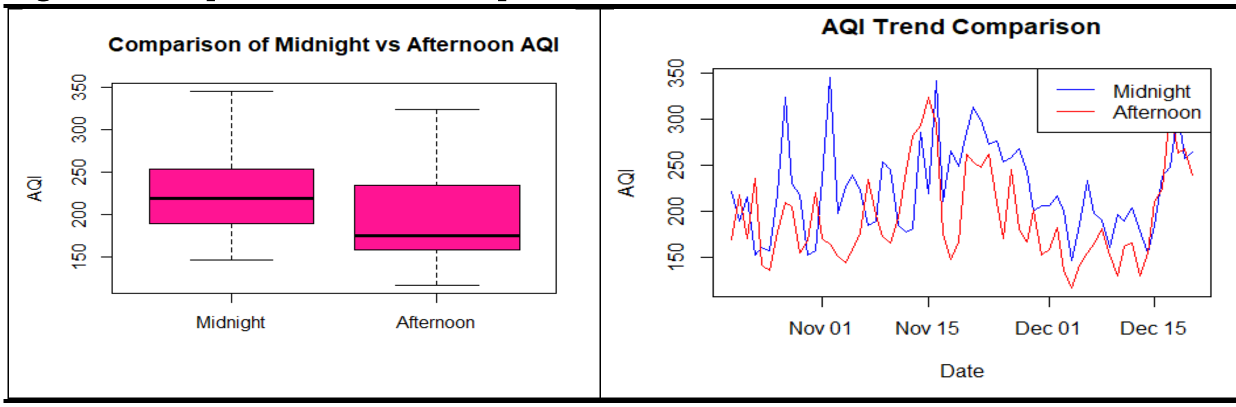
Since inhabitants of Lahore are becoming increasingly concerned about the negative effects of air pollution on their health and well-being, the poor air quality at these hours highlights the urgent need for effective measures to address air pollution in Lahore.

Analysis

Descriptive statistics, Multiple linear regression model and different machine learning modes are bieng applied for further analysis of both the midnight and afternoon data.

A correlation matrix is also bieng made which tells us about the significance of different pollutants in Lahore's AQI.

Figure 1: Comparison on both time period of data



From figure 1 key insights into daily variations can be gained by visually comparing Air Quality Index (AQI) data between midnight and afternoon. The AQI is typically a little higher at midnight than it is in the afternoon, as the boxplot on the left illustrates. A narrower range and marginally lower median values are seen in the afternoon AQI, whereas the midnight AQI has a wider range, with some values exceeding 300, signifying worse air quality. This indicates that, on average, the air quality is better in the afternoon than it is at midnight.

The AQI trends from November to mid-December are depicted in the line graph on the right. Similar erratic patterns with recurring peaks and troughs can be seen in the AQI values at midnight (blue line) and afternoon (red line). Interestingly, over the course of the study, the midnight AQI continuously trends marginally higher than the afternoon AQI, supporting the finding that nighttime air quality is typically poorer. This analysis reveals temporal and diurnal fluctuations in AQI, which may be caused by weather, industrial activity, and traffic.

Table 1: Descriptive statistics for pollutants affecting the AQI (Afternoon, Midnight)

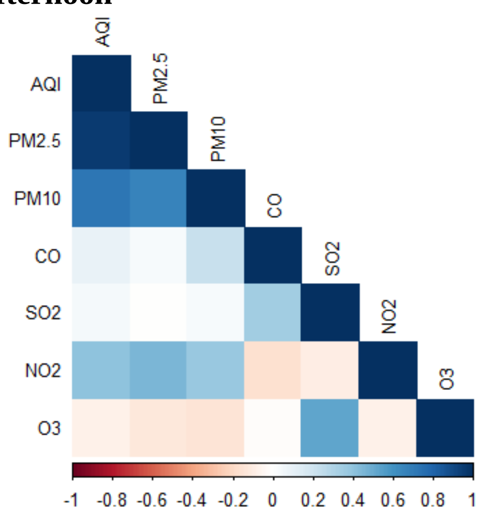
	Afternoon					Midnight				
Pollutants	**N**	**Max**	**Min**	**Mean**	**St. Dev.**	**N**	**Max**	**Min**	**Mean**	**St. Dev.**
AQI	62	324	117	194.47	49.93	62	345	147	223.90	48.40
PM2.5	62	188	50	89.87	27.44	62	191	49	102.76	27.26
PM10	62	251	111	170.48	32.23	62	330	105	194.05	46.07
Co	62	1506	514	937.79	184.35	62	2299	245	913.24	277.89
So2	62	10	3	5.45	1.39	62	13	3	5.58	1.66
No2	62	63	7	14.55	9.35	62	27	6	13.40	3.91
O3	62	22	10	16.16	3.76	62	20	9	14.31	3.41

In table 1 the afternoon AQI ranged from 117 to 324, with an average of 194.47 and a standard deviation (St. d.) of 49.93. PM2.5 varied from 50 to 188, with a mean of 89.87 (St. d: 27.44). With a range of 111 to 251, PM10 displayed a higher average (170.48) and a standard deviation of 32.23. CO levels ranged from 514 to 1506, with a mean of 937.79 and a standard deviation of 184.35. There was little variation in the mean values of SO2 and NO2, which were lower at 5.45 and 14.55, respectively. O3 had a standard deviation of 3.76 and an average of 16.

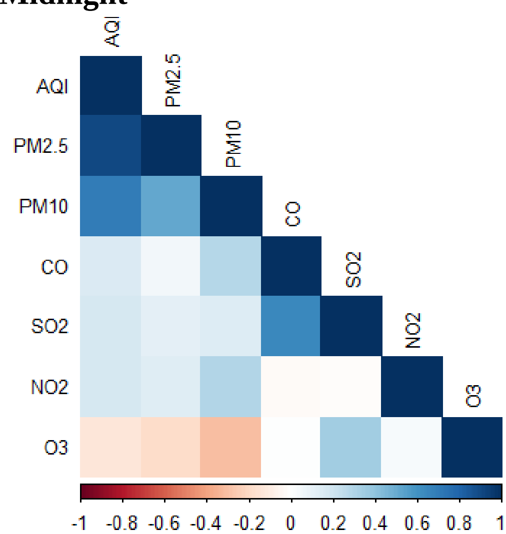
In table 1 the AQI ranged from 147 to 345 at midnight, with an average of 223.90 (St. d: 48.40). While PM10 had a higher mean of 194.05 with more variability (St. d: 46.07), PM2.5 levels also increased somewhat (mean: 102.76, St. d: 27.26). With a mean of 913.24 (St. d: 277.29) and a range of 245 to 2299, CO showed notable variation. O3 marginally dropped to an average of 14.31 (St. d: 3.97), while SO2 (mean: 5.58, St. d: 1.66) and NO2 (mean: 13.40, St. d: 3.91) showed patterns similar to the afternoon.

Figure 2: Pearson correlation matrix

Afternoon



Midnight



In figure 2 the correlation matrix of Afternoon shows the correlation coefficients of the Air Quality index and pollutants including PM2.5, PM10, CO, SO2, NO2, and O3. Whereas NO2 and O3 show weak or moderately negative correlations with AQI, PM2.5, PM10, and CO show strong positive correlations. These associations imply that carbon monoxide and particulate matter have a major role in AQI fluctuations.

In figure 2 the correlation matrix of Midnight shows that PM10 and AQI have a high positive association, while CO and NO2 have moderate correlations. O3 and AQI have a negative correlation, indicating that they have an adverse relationship. The smaller associations between PM2.5 and SO2 suggest that they have less of an effect on AQI.

Table 2: Multiple linear regression for pollutants affecting the AQI (Afternoon, Midnight)

Multiple Linear Regression Model

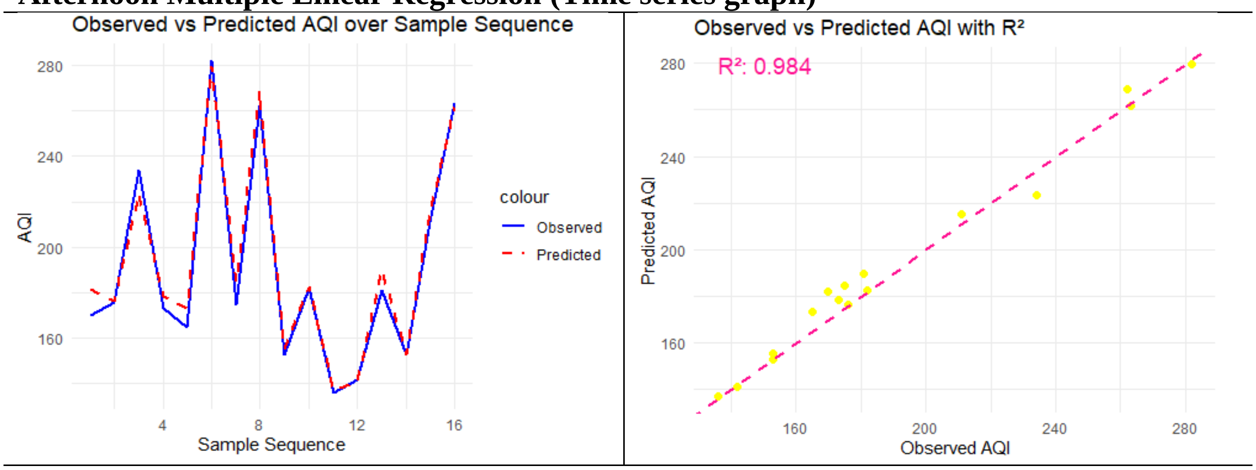
	Afternoon							Midnight						
MLR	**B**	**T**	**R**	**R²**	**VIF**	**F**	**F. Sig**	**B**	**T**	**R**	**R²**	**VIF**	**F**	**F. Sig**
intercept	194.4	191.5			-			223.9	152			-		
PM2.5	28.10	1.83			2			-10.6	-0.4			1.8		
P10	45.30	2.80			1.7			-86.2	-2.7			2.5		
Co	15.7	1.49			1.3			-40.6	-0.9			4.9		
So2	10.78	-0.97	0.99	0.98	1.3	97.6	0.0	42.5	0.4	0.98	0.97	5.9	28.7	0.0
No2	29.48	-2.60			1.7			45.8	1.8			1.9		
O3	10.35	-1.05			1.3			-13.1	-0.6			1.6		

In table 2 the Multiple Linear Regression (MLR) model for Afternoon shows a strong match, accounting for 98.4% of the variation in the AQI ($R2=0.984$). The following predictors are significant: NO2 ($B = -29.48$, $p<0.05$) and PM10 ($B=45.30$, $p<0.05$); CO, SO2, O3, and PM2.5 are not statistically significant. Due to the low Variance Inflation Factors (VIF), multicollinearity is not present. The overall significance of the model is strong ($F=97.65$, $p<0.05$).

In table 2 the (MLR) model for Midnight exhibits a strong fit, accounting for 97.6% of the variance in AQI ($R2=0.976R\ 2=0.976$). Among the factors that are not statistically significant are PM2.5, CO, SO2, and O3, while PM10 ($B= -86.2$, $p<0.05$) and NO2 ($B=45.8$, $p<0.05$) are significant predictors. Moderate multicollinearity is indicated by VIF values, particularly for CO and SO2. Overall, the model is significant ($F=28.75$, $p<0.05$).

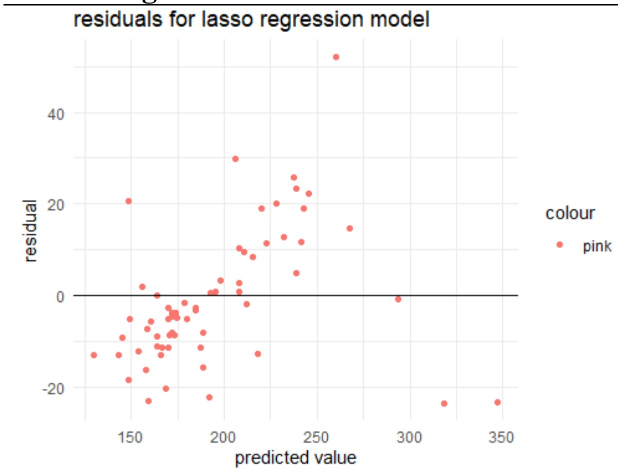
Figure 3: Multiple Linear Regression ND Machine Learning Model (Afternoon)

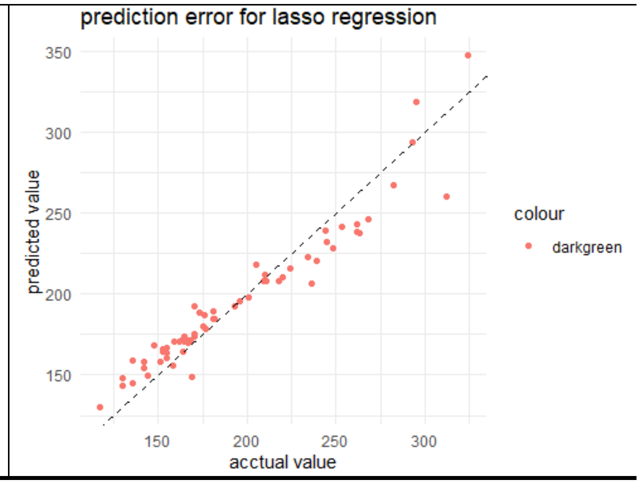
Afternoon Multiple Linear Regression (Time series graph)



Machine Learning Methods

Lasso Regression



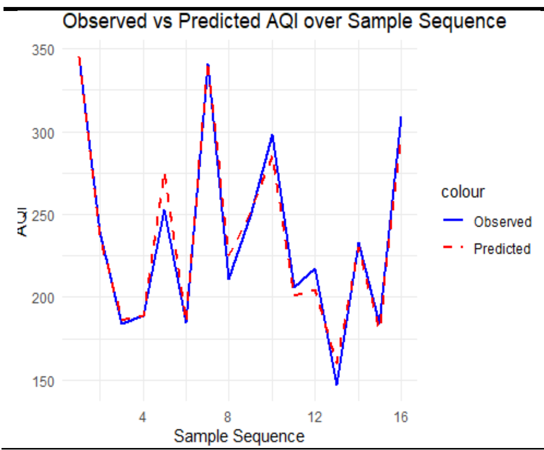


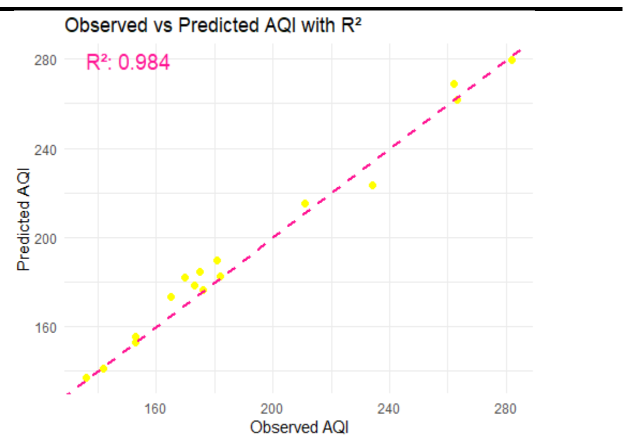
In figure 3 The time series graphic indicates that the observed and predicted AQI values are closely aligned, indicating a high predictive accuracy for the Multiple Linear Regression (MLR) model. The scatter plot, which shows that the predicted values correspond closely along the diagonal line, further supports the model's strong fit with □^{2}=0.984. The Lasso regression model's residuals, which are dispersed around zero, exhibit no discernible pattern or bias in the predictions. A moderate fit is indicated by the prediction error scatter plot, where the anticipated and actual values approximately match, albeit with somewhat greater deviations than MLR. This shows that Lasso regression offers a prediction that is more regularized but marginally less accurate.

Figure 4: Multiple Linear Regression ND Machine Learning Model (Midnight)

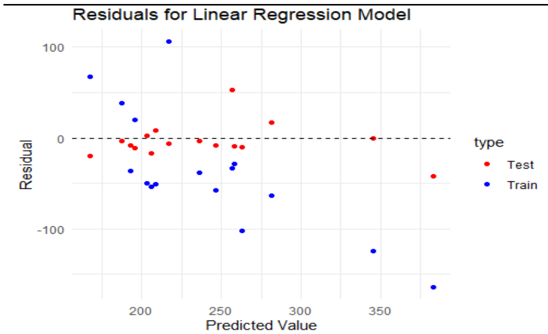
Midnight

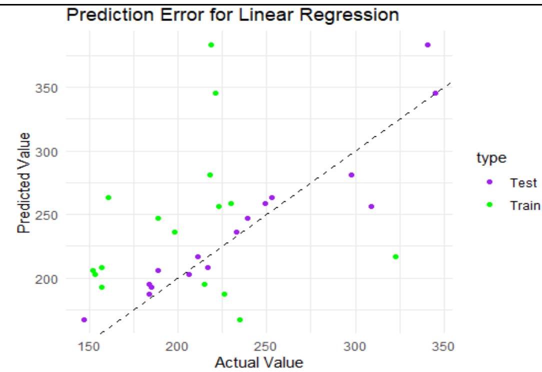
Multiple Linear Regression (Time series graph)





Machine Learning Methods
Linear Regression





In figure 4 the Multiple Linear Regression (MLR) model's findings show that the observed and predicted AQI values fit each other quite well. The first graphic indicates that the model successfully captures the underlying trend because the projected values closely match the observed AQI across the sample sequence. This is further supported by the second plot, which shows that the model accounts for 98.4% of the variance in AQI with a R^2 value of 0.984.
The residual plot for the linear regression model demonstrates that the residuals are uniformly distributed around 0 with no discernible trend, so verifying that the model satisfies important presumptions. In contrast to training residuals, test residuals show a little bit more unpredictability. Most of the points on the plot of prediction error lie along the diagonal line, meaning that the predicted and actual values are aligned well. The model performs about the same on the training and testing data, indicating good generalization.

Table 3: MSE, MAE, MAPE, R^2, RMSE, RMSLE of Machine Learning Models (Afternoon, Midnight)

Machine Learning Models										
Afternoon						**Midnight**				
Models	**MSE**	**MAE**	**MAPE**	**R^2**	**RMSE**	**MSE**	**MAE**	**MAPE**	**R^2**	**RMSE**
LR	218.1	12.1	6.617	0.95	14.77	380.57	13.7	5.6	0.88	19.50
DT	239.1	11.4	5.708	0.87	15.46	719.4	20.8	8.4	0.84	26.8
RF	246.8	12.8	7.275	0.93	15.71	1335.8	24.6	9.3	0.7	36.54
SVM	571.0	17.1	9.947	0.76	23.89	1558.1	21.0	7.2	0.6	39.4
KNN	1567.2	32.0	16.83	0.24	39.58	1704.2	27.5	10.5	0.6	41.2
GB	392.0	14.5	7.901	0.82	19.79	1577.5	31.7	12.6	0.5	39.7
LR	208.6	12.6	6.889	0.97	14.44	388.7	13.9	5.8	0.88	19.7
RR	305.4	15.3	8.499	0.94	17.47	386.8	14.0	5.9	0.88	19,6

The above table 3 gives us the results of MSE, MAE, MAPE, R^2, RMSE with different machine learning techniques for afternoon and midnight and we can see that the best ML technique for afternoon is Lasso Regression as it has lowest MSE and highest R^2 and for midnight the best ML technique is Linear Regression as it has lowest MSE and highest R^2.

Classification

When the response variable becomes qualitative, we must approach classification because this process produces predictions of qualitative responses. When we assign observations into categories through class prediction, we call it classification because it turns the qualitative observation into a classification process. Here in our dataset, we had the qualitative variable too (i.e. Good (0-50), moderate (50-100), poor (100-200), unhealthy (200-300), severe (300-400), hazardous (400-500+).

Figure 5: Confusion matrix of different Machine Learning Model (Midnight)

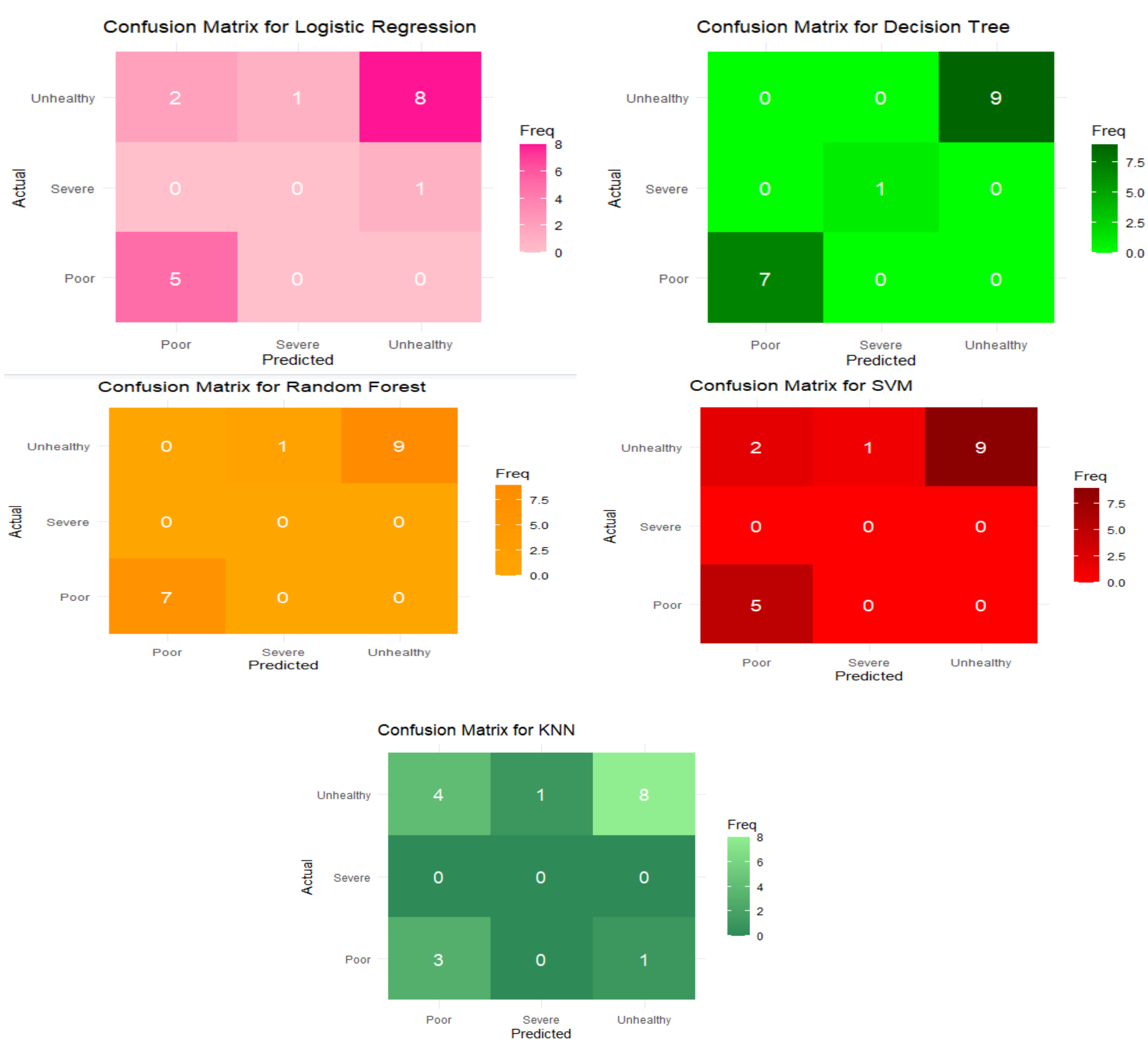


Figure 5 shows the performance of five machine learning models (Logistic Regression, Decision Tree, Random Forest, SVM, and KNN) in classifying the levels of air quality ("Unhealthy," "Severe," and "Poor") is compared based on the confusion matrices.

The logistic regression model is performing at a mediocre level, with considerable misclassifications between "Poor" and "Unhealthy." With the lowest mistakes (mostly green matrix), the Decision Tree model performs the best, which implies that it is the most accurate model in this comparison. Orange matrix for Random Forest model suggests that it performs well too with comparatively less errors.

In contrast, the SVM model performs very poorly, misclassifying most instances (mostly red matrix). The KNN model performs erratically, correctly classifying "Unhealthy" but making glaring misclassifications, especially in the "Poor" category, where the majority are incorrectly categorized as “unhealthy”. In conclusion the decision tree is the most reliable one as compared to the other models.

Figure 6: Boxplot of pollutants affecting the environment by their AQI status (Midnight)

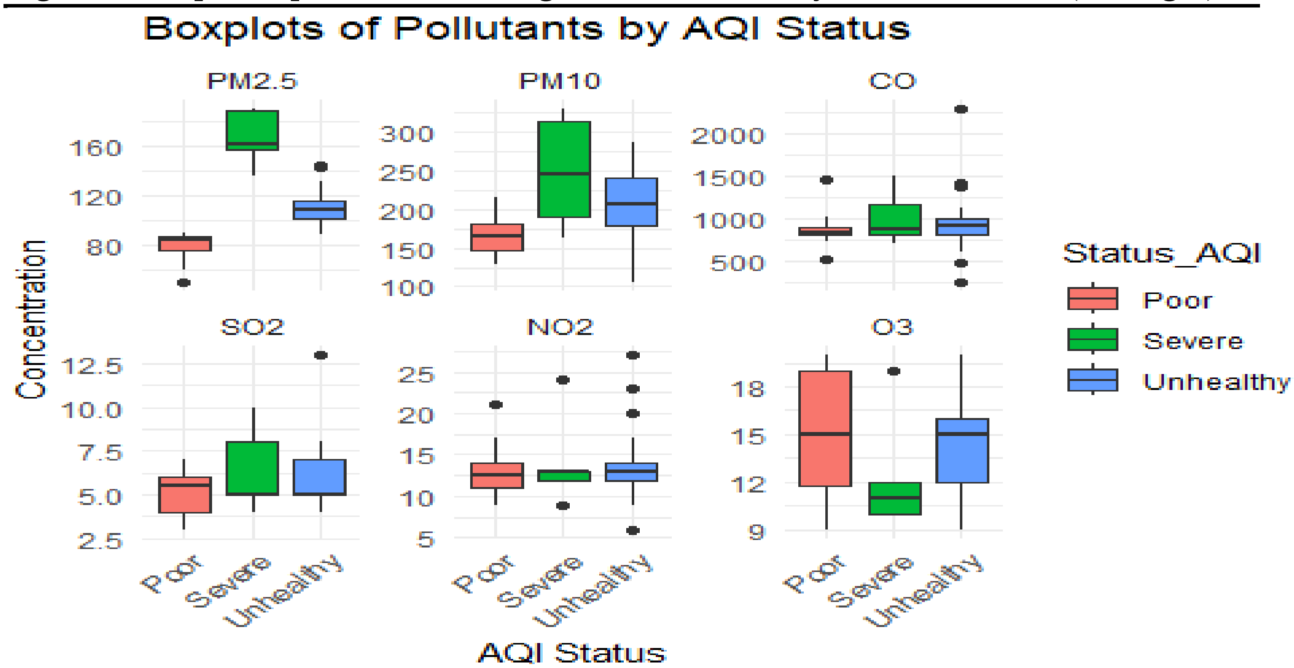


Figure 6 shows the boxplots show how pollutant concentrations vary among the Poor, Severe, and Unhealthy AQI categories. "Severe" conditions result in higher medians and larger distributions of PM2.5 and PM10 levels, which rise noticeably. All categories still have comparatively high CO, although "Severe" circumstances exhibit more variation. Despite their compact distributions, SO2 and NO2 show modest increases in "Severe" and "Unhealthy" situations. The inverse trend of O3 concentrations, which are highest in "Poor" air quality and much lower in "Severe" conditions, suggests that the processes of pollution generation are different. Different pollutant behaviors under different AQI statuses are highlighted by the results.

Figure 7: Boxplot of pollutants by their AQI status (Afternoon)

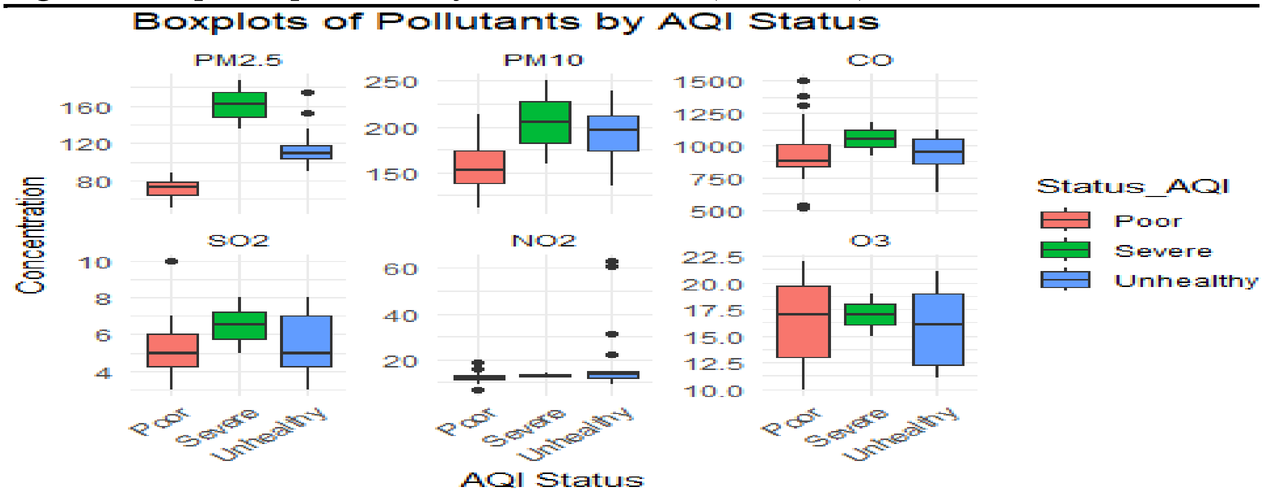
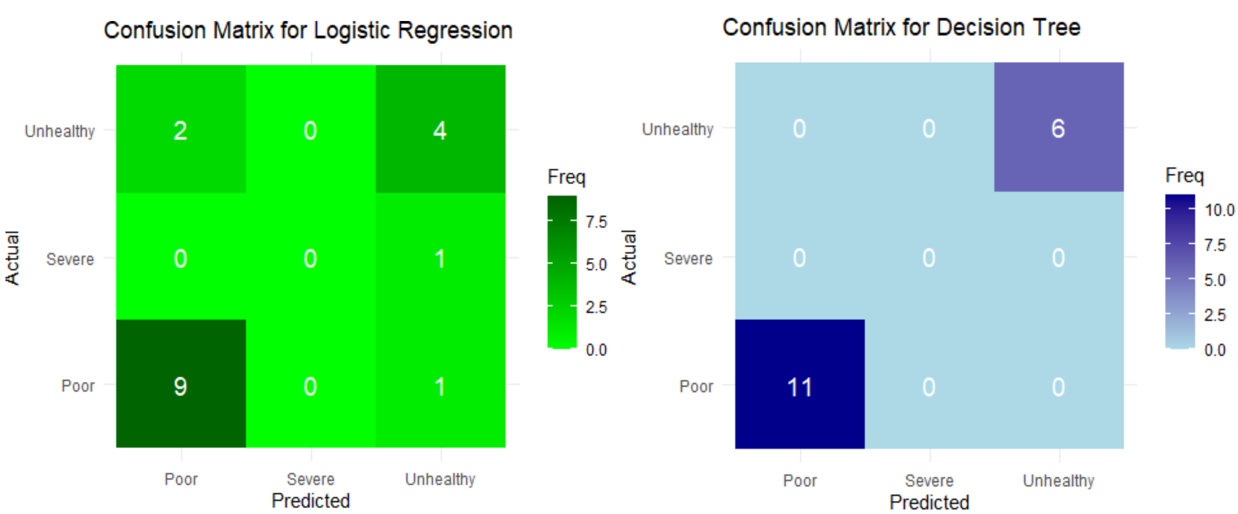


Figure 7 shows the amounts of six pollutants (PM2.5, PM10, CO, SO_2, NO_2, and O_3) in the afternoon at various AQI statuses (Poor, Severe, and Unhealthy). With their concentration rising sharply as air quality deteriorates, PM2.5 and PM10 exhibit a definite upward trend, suggesting a major impact on AQI decline. Carbon monoxide, or CO, is constantly elevated and peaks in the "Unhealthy" category. With significant fluctuation, SO_2 (Sulphur dioxide) also rises as the AQI deteriorates. However, NO_2 (nitrogen dioxide) stays largely constant at all AQI levels, indicating that it might not be the main cause of the deteriorating air quality. Higher AQI conditions tend to increase the concentration of O_3 (ozone), which exhibits greater fluctuation. In general, the main causes of Lahore's bad air quality seem to be PM2.5, PM10, and CO, however SO_2 and O_3 also contribute, albeit less so.

Figure 8: Confusion matrix of different Machine Learning Model (Afternoon)



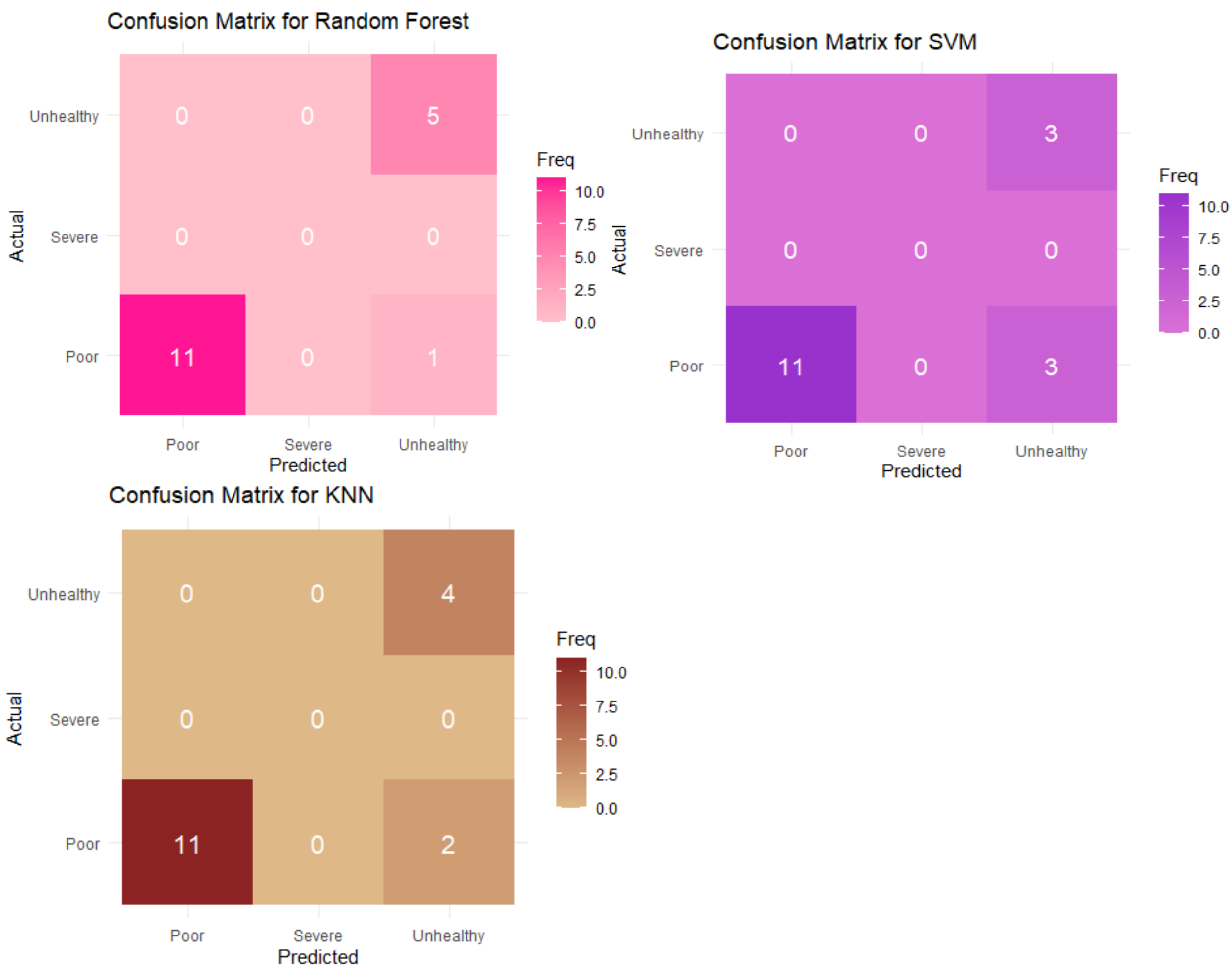


Figure 8 shows how well five distinct machine learning models—Decision Tree, Random Forest, SVM, KNN, and Logistic Regression—perform in dividing data into three groups: Poor, Severe, and Unhealthy. The "Poor" class is consistently well-predicted across all models, but Decision Tree, Random Forest, SVM, and KNN have a strong propensity to properly identify Poor samples (e.g., classifying 11 examples correctly). However, most models are unable to predict any instances in the "Severe" class, indicating weak classification. There are several misclassifications in the "Unhealthy" class, particularly in SVM and Logistic Regression. Interestingly, there is a clearer difference between the Decision Tree and KNN models, indicating higher prediction accuracy. In general, Severe instances are difficult for most models to handle, whereas Poor is the most precisely identified category.

Figure 9: Comparison on different Air pollutants (Left Afternoon, Right Midnight)

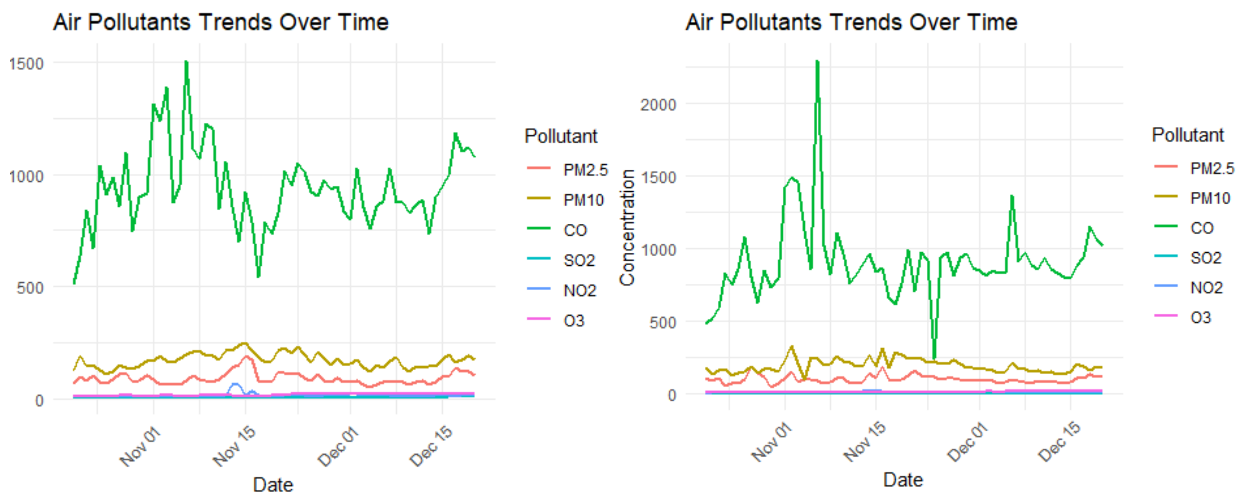


Figure 9 shows the comparison of the afternoon (left) and midnight (right) air pollution trends over time. With discernible variations, PM2.5 and PM10 concentrations are substantially greater than those of other pollutants in both situations. The PM2.5 values in the midnight data show a more noticeable peak, indicating that pollution levels are higher at night. The levels of other pollutants, like CO, SO2, NO2, and O3, are comparatively constant. According to the trends, particulate matter (PM2.5 and PM10) is the most prevalent pollutant. Midnight concentrations are marginally higher, possibly because of increasing emissions from certain sources and decreased air dispersion.

Table 4: Accuracy of different Machine Learning Models (Afternoon, Midnight)

	Midnight					Afternoon				
Model	**Precision**	**F1 Score**	**Recall**	**Accuracy**	**Classification Error**	**Precision**	**F1 Score**	**Recall**	**Accuracy**	**Classification Error**
LR	0.53	0.81	0.57	0.76	0.23	0.74	0.76	0.52	0.76	0.23
DT	1.0	1.0	1.0	1.0	0.00	1.0	1.0	1.0	1.0	0.00
RF	1.0	1.0	1.0	1.0	0.00	0.91	0.93	0.95	0.94	0.05
SVM	0.57	0.84	0.87	0.82	0.17	0.75	0.77	0.89	0.82	0.17
KNN	0.43	0.63	0.68	0.64	0.35	0.83	0.85	0.92	0.88	0.11

Table 4 shows precision, F1 score, recall, accuracy, and classification error of five machine learning models (LR, DT, RF, SVM, and KNN) compared in the table between midnight and afternoon. The two best-performing algorithms are Random Forest (RF) and Decision Tree (DT), with RF marginally declining in the afternoon and DT reaching perfect accuracy (1.0) in both time periods. The accuracy of logistic regression (LR) varies; it is slightly better in the afternoon and poorer at midnight (0.76). While KNN exhibits the biggest classification error and performs the worst, SVM is stable but not as strong as DT or RF. It is better to use DT or RF for deployment and to stay away from KNN.

Discussion

In this paper, we used a set of regression (linear, Lasso, Ridge) and classification (decision tree, random forest, SVM, KNN, gradient boosting) models to analyze pollutant and AQI data collected between October 20 and December 20 2024 to compare the air quality in Lahore in two time points of We aimed to find the most suitable methods of analysis for classification and prediction and the diurnal dynamics of pollution as well. At midnights we observed a generally 15% increase of the levels to the afternoon, where we found that $PM_{2.5}$ and PM_{10} are the main factors of AQI during both periods. We also discovered that Lasso's regularization gave the most stable sort of midnight forecasts while simple linear regression provided the most stable sort of afternoon forecasts. Surprisingly, regarding the intricacy of urban pollution patterns, decision tree and random forest classifiers achieved complete accuracy in the AQI categories classification ("Poor", "Unhealthy", and "Severe") in both periods. The theoretical basis employs a bias-variance trade-off to predict differences in models' performance, photochemical kinetics to explain changes in the secondary pollutants throughout the day, and boundary-layer physics to explain why it is more polluted at night. Besides this, we demonstrate how the pollution trajectory for Lahore represents the early-stage Urbanization patterns by locating our findings in the context of the Environmental Kuznets Curve. This implies that certain nighttime emission rules and particulate filtration arrangements might have the largest impact on the quality of air. These revelations improve our scientific understanding of the phenomenon as they also provide policymakers with workable ideas on how to minimize occasions of peak pollution.

Conclusion

The data was taken at two time periods "afternoon" and "midnight. We have applied the Multiple linear regression and different machine learning models on both of our data sets. The best machine learning model for the afternoon is lasso regression and for Midnight Linear regression works best. We have done classification using machine learning algorithms. The two best-performing algorithms are Random Forest (RF) and Decision Tree (DT), with RF marginally declining in the afternoon and DT reaching perfect accuracy (1.0) in both time periods. We have also done the comparison of different pollutants to see which effect the AQI the most and According to the trends, particulate matter (PM2.5 and PM10) is the most prevalent pollutant. Lastly, we did the comparison of both Midnight and Afternoon and by the result we concluded that AQI has the worst quality at midnight than that at Afternoon.

References

- Al-Eidi, S., Amsaad, F., Darwish, O., Tashtoush, Y., Alqahtani, A., & Niveshitha, N. (2023). Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques. *IEEE Access*, *11*, 115140–115149. https://doi.org/10.1109/ACCESS.2023.3323447
- Amer, A., Mushtaq, N., Albalawi, O. (2024). Forecasting mortality and DALYs from air pollution in SAARC nations. *Scientific Reports, 14,* 25898 https://doi.org/10.1038/s41598-024-76760-9 .
- Choi, Y., Kang, B., & Kim, D. (2024). Utilizing Machine Learning-based Classification Models for Tracking Air Pollution Sources: A Case Study in Korea. *Aerosol and Air Quality Research*, *24*(7). https://doi.org/10.4209/aaqr.230222
- Ct, M., & Amutha, J. (2024). Forecasting of Air Quality Index Using Machine Learning Techniques for Chennai City. In *International Journal of Research Publication and Reviews Journal homepage: www.ijrpr.com* (Vol. 5). www.ijrpr.com

- 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*, *2023*, 1–26. https://doi.org/10.1155/2023/4916267
- Lei, T. M. T., Siu, S. W. I., Monjardino, J., Mendes, L., & Ferreira, F. (2022). Using Machine Learning Methods to Forecast Air Quality: A Case Study in Macao. *Atmosphere*, *13*(9). https://doi.org/10.3390/atmos13091412
- Liu, Q., Cui, B., & Liu, Z. (2024). Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling. *Atmosphere*, *15*(5). https://doi.org/10.3390/atmos15050553
- Loganathan, A., Sumithra, P., & Deneshkumar, V. (2022). Estimation of Air Quality Index Using Multiple Linear Regression. *Applied Ecology and Environmental Sciences*, *10*(12), 717–722. https://doi.org/10.12691/aees-10-12-3
- Mampitiya, L., Rathnayake, N., Leon, L. P., Mandala, V., Azamathulla, H. M., Shelton, S., Hoshino, Y., & Rathnayake, U. (2023). Machine Learning Techniques to Predict the Air Quality Using Meteorological Data in Two Urban Areas in Sri Lanka. *Environments - MDPI*, *10*(8). https://doi.org/10.3390/environments10080141
- Nataraj, R., & Goldsztein, G. (n.d.). *Application of Machine Learning on Air Quality*. www.JSR.org
- Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-54807-1
- Rahman, M. M., Nayeem, M. E. H., Ahmed, M. S., Tanha, K. A., Sakib, M. S. A., Uddin, K. M. M., & Babu, H. M. H. (2024). AirNet: predictive machine learning model for air quality forecasting using web interface. *Environmental Systems Research*, *13*(1). https://doi.org/10.1186/s40068-024-00378-z
- Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam. *Chemosphere*, *338*. https://doi.org/10.1016/j.chemosphere.2023.139518
- Razzaq, A., Zafar, M. M., Zahra, L. T., Qadir, F., Qiao, F., Ullah, M. H., Shehzad, S., Rasool, G., & Jiang, X. (2024). Smog: Lahore needs global attention to fix it. In *Environmental Challenges* (Vol. 16). Elsevier B.V. https://doi.org/10.1016/j.envc.2024.100999
- Suroshe, S., Sanjay Suroshe, S., Student, P., Dharpal, S., & Ingole, N. W. (n.d.). Prediction of Air Quality Index Using Regression Models. *GIS Science journal*, *9*, 2022. https://www.researchgate.net/publication/369763189