Air Quality Prediction and Analysis - Team 9

Pranav Kapparad - 211AI026
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: pranavkapparad30@gmail.com

Satyam Agrawal - 211AI044
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: lmcsatyam@gmail.com

Abstract—**Air pollution is one of the most pressing issues in our times necessitating immediate attention. Air quality is crucial for both our health individually and environmentally. Hence, accurately predicting air quality is of paramount importance to manage the degradation and work improving air quality.This paper focuses on predicting air quality during the COVID-19 pandemic lockdown by designing and testing a model using a dataset collected in Wuhan, China. The dataset consists of 488 time-series data points from September 2019 to December 2020, encompassing several pertinent parameters including Air Quality Index (AQI). The methodology section outlines the various techniques applied to the dataset. The missing values were handled using the K Nearest Neighbors (KNN) imputer. Additionally, exploratory data analysis techniques were employed, including scatter plots, time series graphs, correlation matrix, and covariance matrix analysis, to gain insights into the relationships between variables.Classification and regression techniques were utilised for air quality prediction. For classification, the AQI values were categorised into different classes representing the severity of pollution. Classification models were applied with and without feature selection, as well as with and without the Synthetic Minority Over-sampling Technique (SMOTE) to address imbalanced classes. Regression analysis was conducted on both the original and residual data obtained through additive time series analysis.The experimental results divulged the effectiveness of SMOTE in improving the accuracy of classification models. Decision Tree, Random Forest, and Gradient Boosting exhibited superior performance among the classification models. Regression analysis demonstrated that Random Forest Regression and Polynomial Regression yielded the best results for predicting AQI values. In essence, this study provides valuable insights into the prediction of air quality during the COVID-19 pandemic lockdown. The findings contribute to the understanding of the impact of lockdown measures on air pollution and emphasise the importance of implementing effective measures to manage pollution. The model can inform policymakers for proactive measures to improve public health and environment.**

I. Introduction

The quality of air we breathe is crucial for the health of individuals and societies. Poor air quality exposes us to various respiratory ailments such as asthma, bronchitis, lung cancer, and chronic obstructive pulmonary disease (COPD). Long-term exposure to polluted air is also linked with cardiovascular diseases. Children exposed to poor air quality are at a higher risk of impaired lung development. Air quality also negatively affects the environment as a whole, leading to climate change, acid rain, ozone depletion, and smog formation. Poor air quality is a threat to both humans and the environment. Hence, it is of paramount importance to prevent air quality from worsening. Therefore, a rational prediction of air quality is valuable for taking necessary measures.

At the end of 2019, a new coronavirus broke out in Wuhan, China, which rapidly transmitted across the world. In order to contain the virus, several governments imposed lockdowns resulting in lesser economic activity. An unintended but welcome consequence of these lockdowns was an improvement in air quality. These results displayed that air quality is affected by both seasonal factors and social factors. The improvement of air quality during the pandemic provides scope for governments to enact new laws to manage pollution such as traffic and industrial restrictions.

Hence, prediction of air quality during the pandemic emanated lockdown is of social importance. We aim to contribute to this cause by designing an accurate model for predicting air quality. The model is designed and tested using air quality dataset of Wuhan collected during and after the implementation of lockdown. The paper discusses both classification and regression techniques in order to predict the AQI values. Methods such as SMOTE and Additive Time Series Analysis are also discussed.

A. Dataset Discussion

The dataset provides an accurate distinction in air quality during and after the lockdown. The Wuhan dataset is a time series dataset with 488 data points ranging from 2019/9/1 to 2020/12/31. The dataset has the following parameters: air quality index (AQI), PM2.5, PM10, SO2, NO2, O3, CO, l-temp, h-temp, wet, wind, hpa, visibility, precipitation, and cloud.

The paper is structured in a way that section II discusses the literature surveyed, the next section talks about the methodology employed in detail. Section IV consists of the experimental results and analysis carried out. The paper ends with Section V which carries the conclusion.

II. Literature Survey

[1]: Paper titled Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare discusses classification of AQI busket values obtained from the metro citites of India. This employed a SVM model with optimised kernel values to predict AQI values. To avoid and reduce the effects of minority classes, SMOTE was used which showed an improvement in the obtained results.

[2]: Paper titled A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic by Zhao et al. was used as the base paper for the analysis carried out. The paper investigated the effects of lockdown on China's air pollution and tried to develop a deep learning model to predict the same. It mentioned the use of separation of seasonal and lockdown components to obtain a residual component, a technique borrowed in this paper. They further used spatial autocorrelation and fetature selection to obtain and compare prediction results.

III. Methodology

In this section, discussion about the plethora of methods and analysis applied on the data is mentioned. The process begins with handling of missing values:

A. Handling Missing Values

The missing values in the dataset were handled using three different methods and compared.

1) **Drop Rows**: The missing values were first handled by dropping all the rows with null values. This led to around 15 percent reduction in the size of the dataset, significant to interfere with the accuracy of the model and hence discarded.
2) **Mean and Median**: The features that were almost symmetrical i.e skewness of magnitude less than 0.5 were imputed using median values and the features with skewness of magnitude greater than 0.5 were imputed using mean values.
3) **K Nearest Neighbour (KNN) imputer**: The process was carried out using KNN Imputer imported from the 'sklearn' module. This imputer utilises k-Nearest Neighbours method to fill in the missing values with the mean of the k nearest neighbours. The distance is, by default, measured using the Euclidean distance metric.

Since the best results were given by KNN imputing, it was employed for handling the missing values.

B. Date Format Manipulation

Given the time series data, the dataset has 'date' as one of the columns. The value in the column is in the format yyyy/mm/dd. In order to better work with the data, the date was split into components i.e separate columns for day, data and year.

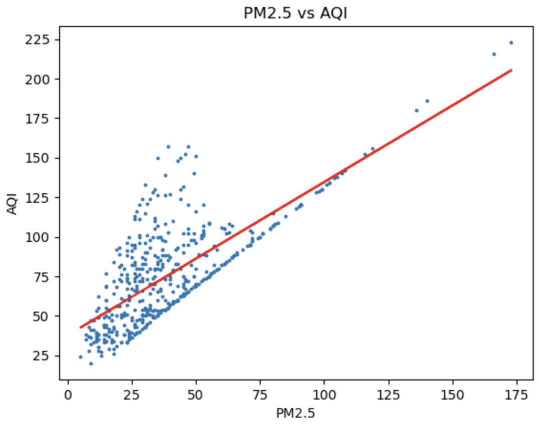
C. Exploratory Data Analysis

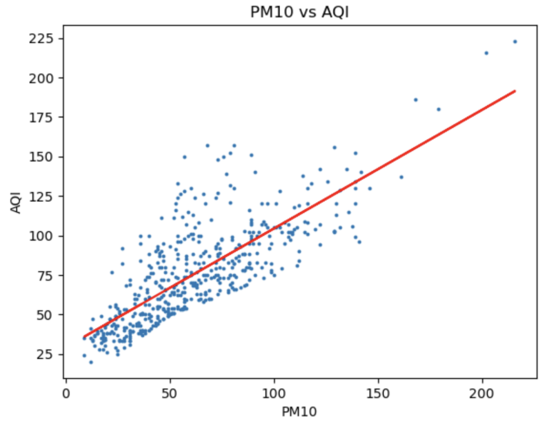
Scatter Plots:

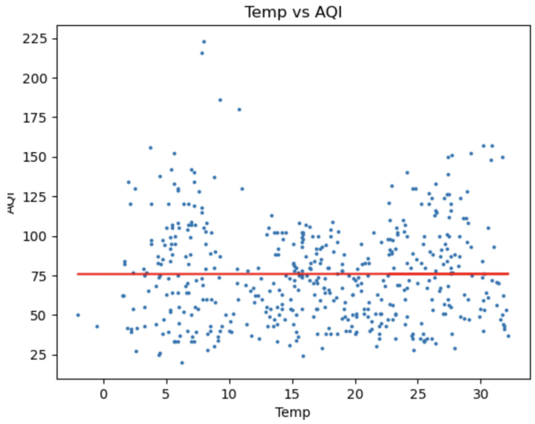
To gain a better understanding of the relationships between different parameters in the dataset, particularly with AQI, several scatter plots were plotted with the best fit line using matplotlib. Scatter plots are an informative type of data visualization that shows relationships between variables.

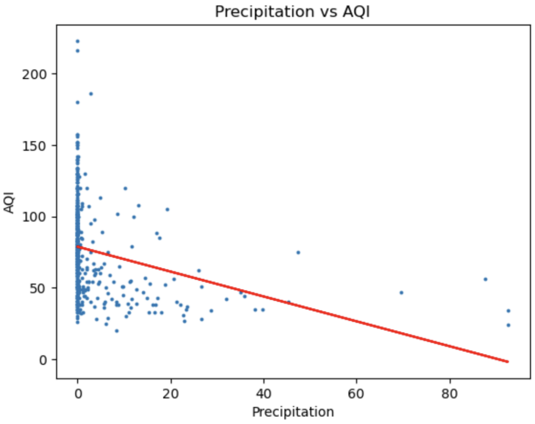
Key Observations:

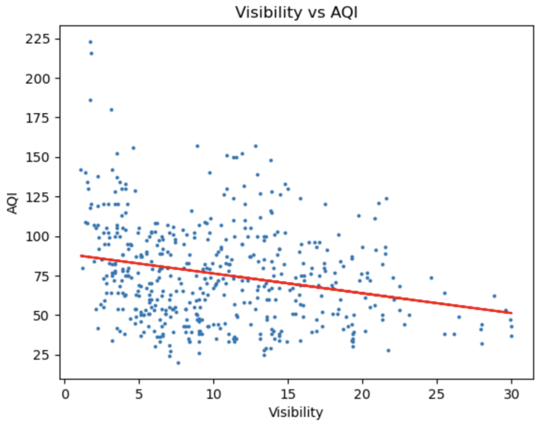
1) Strong positive correlation between AQI and PM10, PM2.5.
2) Insignificant correlation between AQI and temperature.
3) Significant negative correlation between AQI and wind.
4) Positive correlation of AQI with SO2, NO2, CO, and O3.

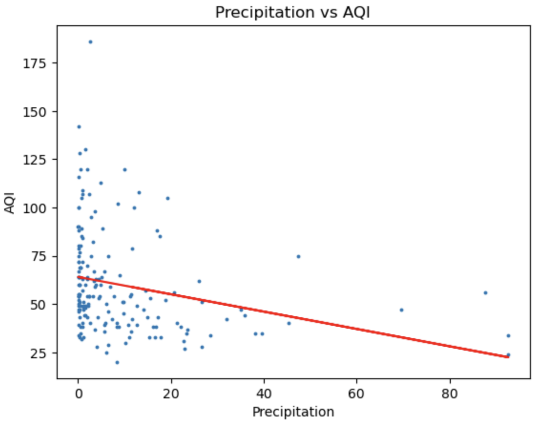


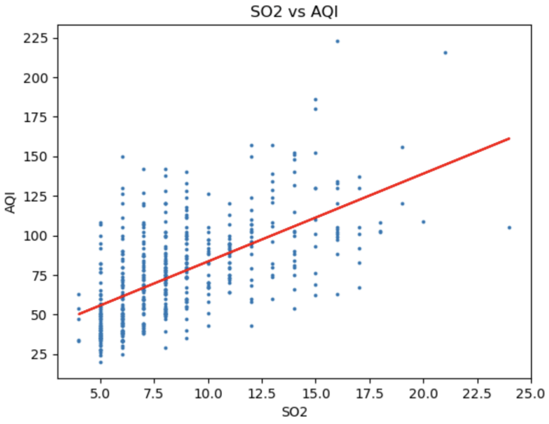


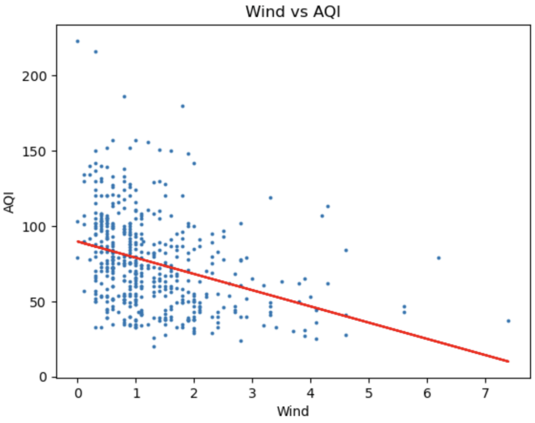


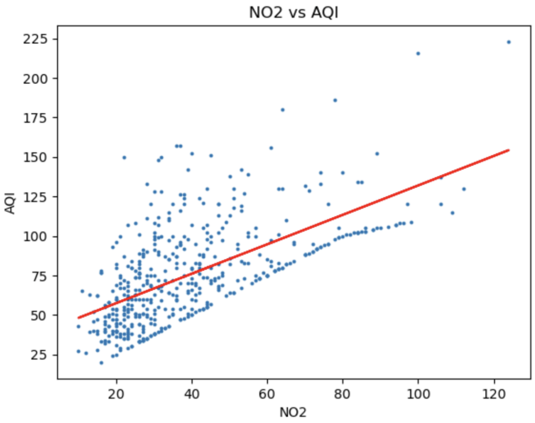

Precipitation vs AQI
AQI
0
50
100
150
200
0
20
40
60
80
Precipitation

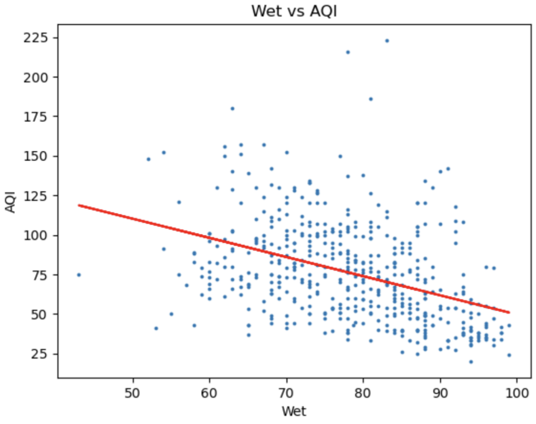

Visibility vs AQI
AQI
25
50
75
100
125
150
175
200
225
0
5
10
15
20
25
30
Visibility

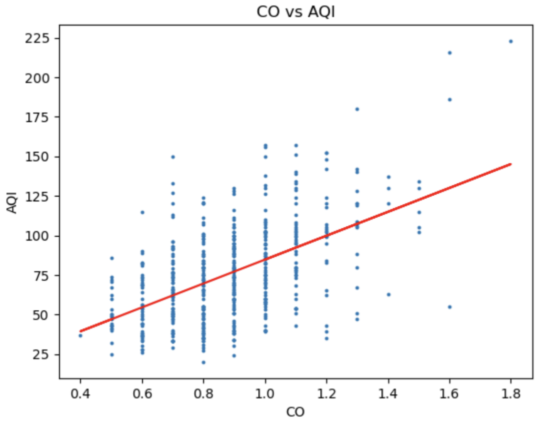

Precipitation vs AQI
AQI
25
50
75
100
125
150
175
0
20
40
60
80
Precipitation

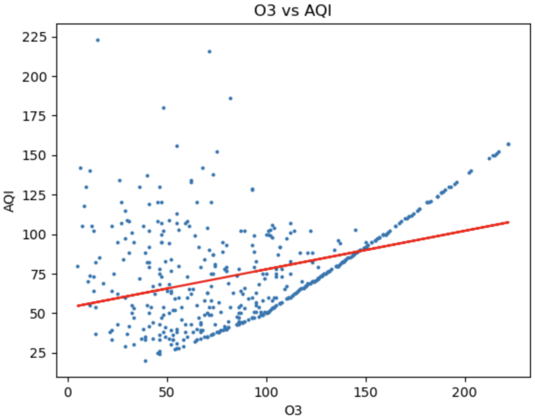

SO2 vs AQI
AQI
25
50
75
100
125
150
175
200
225
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
SO2

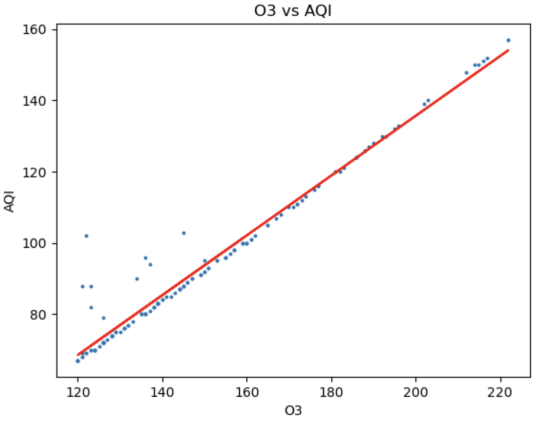

Wind vs AQI
AQI
0
50
100
150
200
0
1
2
3
4
5
6
7
Wind


NO2 vs AQI
AQI
25
50
75
100
125
150
175
200
225
20
40
60
80
100
120
NO2


Wet vs AQI
AQI
25
50
75
100
125
150
175
200
225
50
60
70
80
90
100
Wet

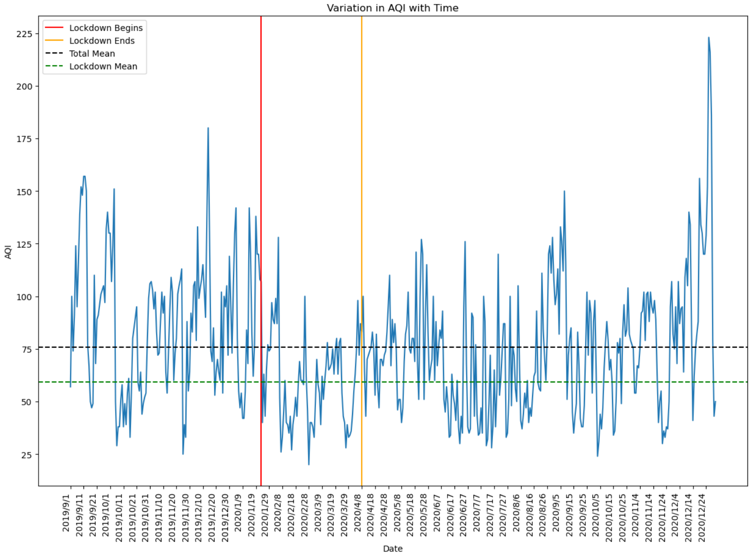

CO vs AQI
AQI
25
50
75
100
125
150
175
200
225
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
CO





Time Series Graphs:

Time series graphs were plotted to visualize the values of AQI over a period of time. Time series graphs are a type of data visualization that represents the values of a parameter over time. Time series graphs of AQI were plotted to visualize the improvement in AQI during the lockdown period. The means of AQI with and without lockdown were calculated to verify the difference. The mean AQI during the lockdown was calculated to be 75.89 compared to 59.14 without lockdown.



D. Correlation Matrix

A correlation matrix is a statistical technique used to evaluate the relationship between variables in a dataset. The correlation between the variables is displayed in the form of an n-by-n matrix, where n is the number of parameters. The correlation matrix of the dataset was plotted and visualized as a heatmap to gain additional insights into the dataset.

Key Observations:

1) Significant positive correlation between O3 and temperature.
2) Significant negative correlation between visibility and CO.
3) Significant negative correlation between SO2 and wet.
4) Significant negative correlation between NO2 and wind.
5) Significant negative correlation between O3 and cloud.

These observations were further explored using scatter plots.

E. Covariance Matrix

A covariance matrix is used to describe the covariance values between the parameters. Similar to the correlation matrix, the covariance matrix is represented as an n-by-n matrix. The covariance matrix of the dataset was plotted and visualized as a heatmap.

F. Classification

Classification involves the bucketing of data points based on their features. The target variable in this study is the AQI value, which was categorized into different classes representing the severity of pollution. The following AQI categories and ranges were used:

- Good (0-50)
- Satisfactory (51-100)
- Moderate (101-200)
- Poor (201-300)
- Very poor (301-400)

- Severe (401-500)

Classification was performed in four stages, following the specified order:

Classification - Without Feature Selection - Without SMOTE: Initially, basic classification models were applied to the data without feature selection or SMOTE. The representative models used were:

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Support Vector Machine
- K-Nearest Neighbors
- Gaussian Naive Bayes
- Multi-Layer Perceptron

Cross-validation with 5 folds was used to test the accuracy of each of the above mentioned models. After repeated tests, it was found that Decision Tree, Random Forest and Gradient Boosting outperformed the others in terms of accuracy score after cross validating.

Classification - Without Feature Selection - With SMOTE: SMOTE stands for Synthetic Minority Over-sampling Technique. The word Synthetic refers to the fact that the data added is data created artificially. Minority refers to the class which is not well represented. A class with more representation usually introduces a bias against those with fewer data points, hence one needs to create more samples of the under represented classes to avoid this problem. This is the basis and reasoning behind applying SMOTE and observing its effects. The models applied were same as those applied without SMOTE and a marked increase in the accuracy obtained after cross-validation was observed.

Classification - With Feature Selection - Without SMOTE: Feature Selection is a technique used to select the most significant/relevant features out of a dataset. This was especially important in case of Air Quality Prediction analysis since it helps develop and understanding of the features that contribute the most towards increasing AQI levels. Several methods were employed to carry this out. They are:

- Univariate Feature Selection - Using Correlation: PM10, PM2.5, NO2, SO2, CO, wet, cloud
- Univariate Feature Selection - Using Mutual Information: PM2.5, PM10, SO2, NO2, O3, CO, cloud
- Using SelectKBest: PM2.5, PM10, SO2, NO2, O3, precipitation, cloud
- Using RFE - Gradient Boost: PM2.5, PM10, NO2, O3, CO, cloud, wind

Some observations made: PM2.5, PM10, NO2, cloud are common to all the four feature selected sets. Next features in order are O3, CO, SO2, wind, precipitation, wet. Based on these observations, two more feature sets were included:

- PM2.5, PM10, NO2, cloud, SO2, O3, precipitation, wind, CO
- PM2.5, PM10, NO2, cloud, SO2, O3, precipitation, wind

Once again, classification was carried out on the mentioned feature sets and the cross validated scores were noted.

Classification - With Feature Selection - With SMOTE: In this stage, SMOTE was applied to the data after performing feature selection. The classification models were then applied to the transformed dataset to evaluate the performance.

G. Regression

Regression is the prediction of continuous numerical target using the feature input variables. In the case of Air Quality Prediction, AQI value is the target feature that we are trying to predict the value for in this paper. Regression was carried out both, in a manner which was temporal as well as in a non-temporal manner. Also, to mitigate the seasonal effects and covid effects, detrending of data was carried out and a residual data was obtained.

Analysis and Prediction on Original Data: To start with, Regression was carried out on the original data which had all the components intact. Models used for the same were:

- Linear Regression
- Random Forest Regression
- Polynomial Regression
- Decision Tree Regression
- Support Vector Regression

In order to evaluate, three metrics were used, namely Mean Absolute Error (MAE), Mean Squared Error (MSE) and R^2 score. A small explanation for each of them is necessary, and is given below:

- **MAE**: Mean absolute error refers to the mean of absolute difference between the predicted and actual values.
- **MSE**: Mean squared error refers to the mean of square of difference between the predicted and actual values.
- R^2 **Score**: R^2 scores helps us understand how well a model has captured the variability of the data. It varies from 0 to 1. A higher value indicates a better model as it represents the fact that model has understood the data variability.

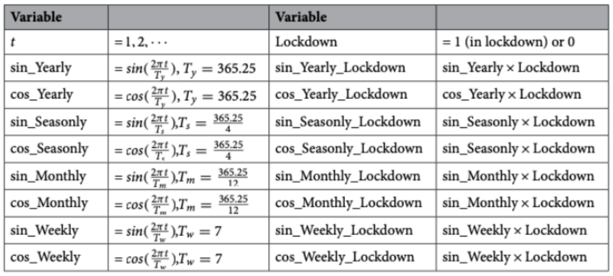
Random Forest Regression and Polynomial Regression were found to be the top performing models on the basis of combined evaluation of the three metric mentioned. Thus, feature selection was applied to both the methods in order to improve the performance and increase the metric scores.

Recursive feature elimination was used to carry out the same and a marked improvement was observed.

Temporal Analysis: Temporal analysis involved splitting the data into training and test sets based on a specific time period. The first 350 days were used as training data, and the remaining days were used as test data. The results of the temporal analysis are discussed in the subsequent sections.

Analysis and Prediction on Residual Data: The original data is now treated to separate out the lockdown and seasonal component using additive time series analysis. The formula used for the same is depicted in the following image:

$$\begin{aligned}S_t = {} & a_0 + a_1 t + a_2\text{sin\_Yearly} + a_3\text{cos\_Yearly} + a_4\text{sin\_Seasonly} + a_5\text{cos\_Seasonly}\\ & + a_6\text{sin\_Monthly} + a_7\text{cos\_Monthly} + a_8\text{sin\_Weekly} + a_9\text{cos\_Weekly}\\ & + a_{10}\text{Lockdown} + a_{11}\text{sin\_Yearly\_Lockdown} + a_{12}\text{cos\_Yearly\_Lockdown}\\ & + a_{13}\text{sin\_Seasonly\_Lockdown} + a_{14}\text{cos\_Seasonly\_Lockdown}\\ & + a_{15}\text{sin\_Monthly\_Lockdown} + a_{16}\text{cos\_Monthly\_Lockdown}\\ & + a_{17}\text{sin\_Weekly\_Lockdown} + a_{18}\text{cos\_Weekly\_Lockdown},\end{aligned}$$



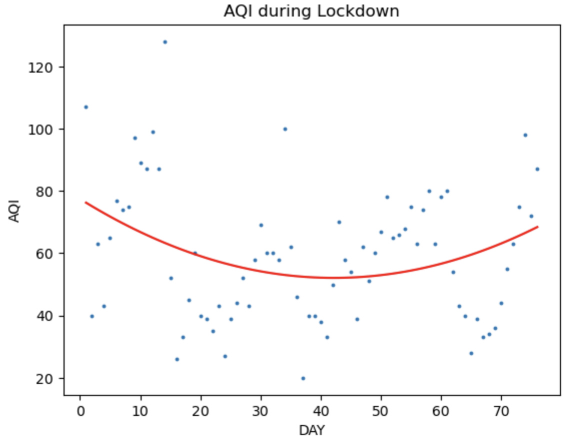
Variable		Variable	
t	$= 1, 2, \cdots$	Lockdown	= 1 (in lockdown) or 0
sin_Yearly	$= sin(\frac{2\pi t}{T_y}), T_y = 365.25$	sin_Yearly_Lockdown	sin_Yearly × Lockdown
cos_Yearly	$= cos(\frac{2\pi t}{T_y}), T_y = 365.25$	cos_Yearly_Lockdown	cos_Yearly × Lockdown
sin_Seasonly	$= sin(\frac{2\pi t}{T_s}), T_s = \frac{365.25}{4}$	sin_Seasonly_Lockdown	sin_Seasonly × Lockdown
cos_Seasonly	$= cos(\frac{2\pi t}{T_s}), T_s = \frac{365.25}{4}$	cos_Seasonly_Lockdown	sin_Seasonly × Lockdown
sin_Monthly	$= sin(\frac{2\pi t}{T_m}), T_m = \frac{365.25}{12}$	sin_Monthly_Lockdown	sin_Monthly × Lockdown
cos_Monthly	$= cos(\frac{2\pi t}{T_m}), T_m = \frac{365.25}{12}$	cos_Monthly_Lockdown	sin_Monthly × Lockdown
sin_Weekly	$= sin(\frac{2\pi t}{T_w}), T_w = 7$	sin_Weekly_Lockdown	sin_Weekly × Lockdown
cos_Weekly	$= cos(\frac{2\pi t}{T_w}), T_w = 7$	cos_Weekly_Lockdown	sin_Weekly × Lockdown

To obtain the coefficients, Ordinary Least Square method was used due to its simplicity and speed. After obtaining the residual data, all the analysis/prediction performed on the original data is replicated for the residual data. The results obtained are discussed in the next section.

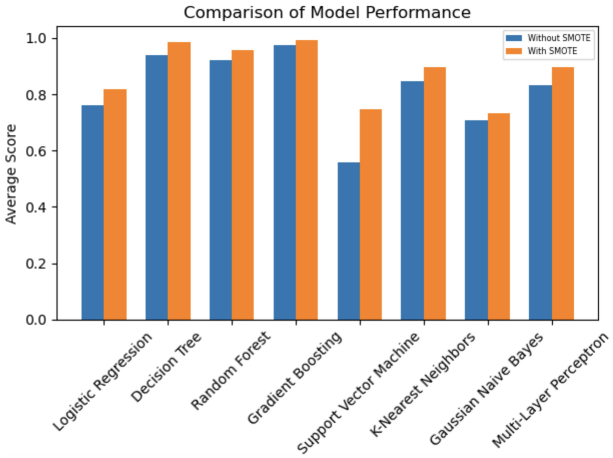
IV. Experimental Results and Analysis

1) Some empirical analysis on the mean of AQI and PM2.5:

1) There is a general trend of reduction in the mean of pollutant indicators during lockdown, which is completely justified given the reduction in polluting factors.
2) The reduction in PM2.5 levels was not too evident and required further exploration.
3) On further investigation, it was found that the minimum level of PM2.5 was recorded between the 21st and 41st day of the lockdown. This lower bound was in sync with the AQI level, a testimony to their strong correlation.
4) One possible explanation could be that people were the most compliant with the lockdown measures during this specific period. However, this explanation seems less plausible as people are expected to be more fearful and compliant at the beginning rather than in the middle. Additionally, this explanation assumes that PM2.5 levels quickly return to a ”normal” condition.
5) A more reasonable explanation would be that it takes time for PM2.5 levels to settle down, and hence the greatest effect is seen between the 21st and 41st days of the lockdown.
6) This also highlights the fact that people gradually became more lenient with the rules over time.
7) A degree 2 regression on the pollutants can help visualize and test the hypothesis.

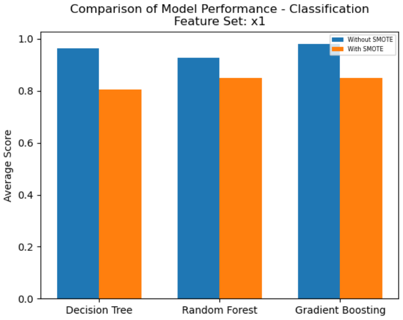


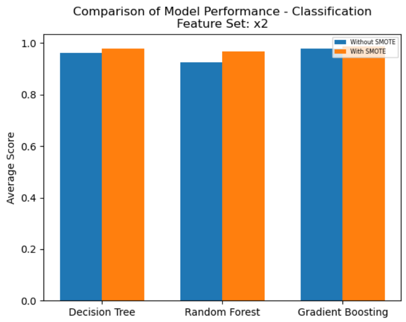
2) Results for Classification without feature selection:

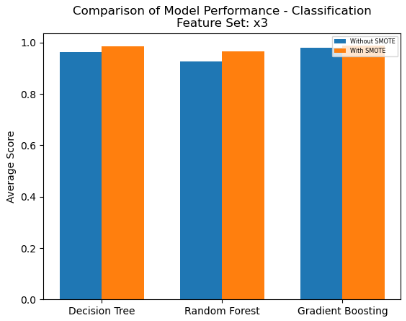


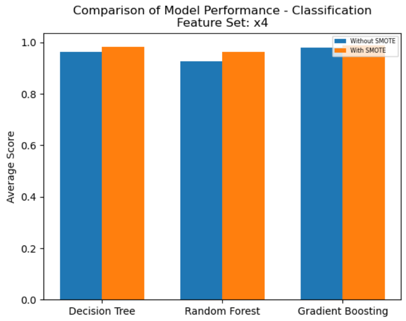
It is evident that classification with SMOTE provide better results than one without SMOTE for every model. It is also noted that Decision Tree, Random Forest and Gradient Boosting are the top 3 models in terms of accuracy.

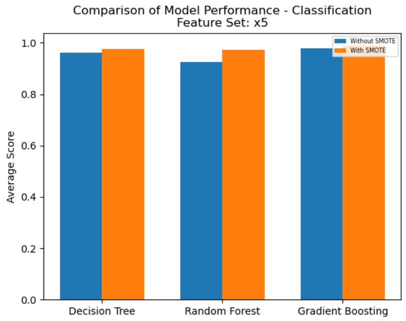
3) Results for Classification with feature selection:

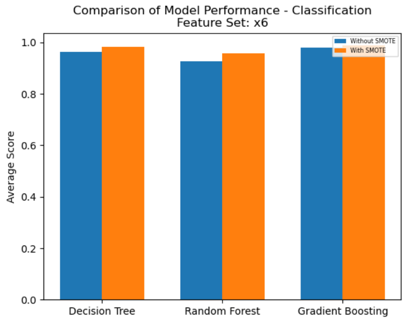






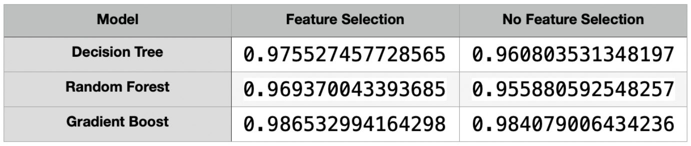






As it has been the case, SMOTE outperforms the Non-SMOTE classification except for the feature set x1. It was also observed that set x3 produced the best results consistently. x3 is composed of: PM2.5, PM10, SO2, NO2, O3, precipitation, cloud

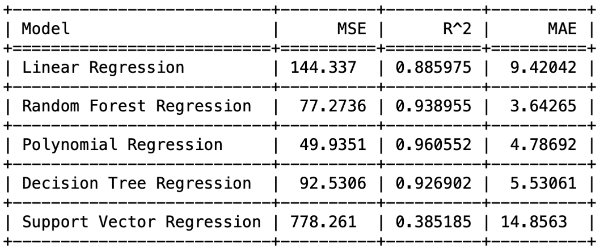
4) Analysing the effectiveness of Feature Selection: The graph below shows the comparison between best performing feature selected result and the complete dataset:



Model	Feature Selection	No Feature Selection
Decision Tree	0.975527457728565	0.960803531348197
Random Forest	0.969370043393685	0.955880592548257
Gradient Boost	0.986532994164298	0.984079006434236

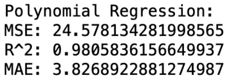
One can observe that the difference is not very significant. This might be due to the fact that original data does not have too many features.

5) Results for Regression on Original Data:



```
+--------------------------+----------+----------+----------+
| Model                    |      MSE |      R^2 |      MAE |
+==========================+==========+==========+==========+
| Linear Regression        | 144.337  | 0.885975 |  9.42042 |
+--------------------------+----------+----------+----------+
| Random Forest Regression |  77.2736 | 0.938955 |  3.64265 |
+--------------------------+----------+----------+----------+
| Polynomial Regression    |  49.9351 | 0.960552 |  4.78692 |
+--------------------------+----------+----------+----------+
| Decision Tree Regression |  92.5306 | 0.926902 |  5.53061 |
+--------------------------+----------+----------+----------+
| Support Vector Regression| 778.261  | 0.385185 | 14.8563  |
+--------------------------+----------+----------+----------+
```

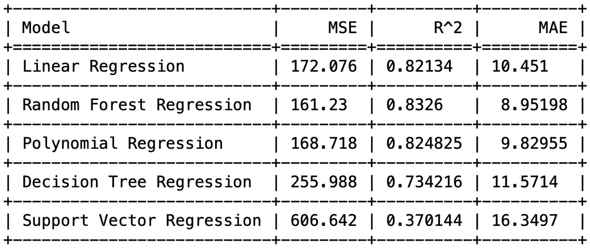
It is clear from observation that Polynomial Regression is the best performing model. Hence a feature selection using RFE was carried out which further improved the accuracy. The details are:



```
Polynomial Regression:
MSE: 24.578134281998565
R^2: 0.9805836156649937
MAE: 3.8268922881274987
```

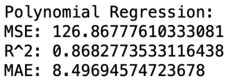
Fig. 1: Improved result after feature selection

6) Results for Regression on Residual Data:



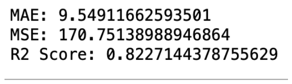
```
+--------------------------+---------+----------+----------+
| Model                    |     MSE |      R^2 |      MAE |
+==========================+=========+==========+==========+
| Linear Regression        | 172.076 | 0.82134  | 10.451   |
+--------------------------+---------+----------+----------+
| Random Forest Regression | 161.23  | 0.8326   |  8.95198 |
+--------------------------+---------+----------+----------+
| Polynomial Regression    | 168.718 | 0.824825 |  9.82955 |
+--------------------------+---------+----------+----------+
| Decision Tree Regression | 255.988 | 0.734216 | 11.5714  |
+--------------------------+---------+----------+----------+
| Support Vector Regression| 606.642 | 0.370144 | 16.3497  |
+--------------------------+---------+----------+----------+
```

Since Polynomial Regression and Random Forest, both of them performed well, feature selection was applied to the two in order to take a decision, the results for the same are:



```
Polynomial Regression:
MSE: 126.86777610333081
R^2: 0.8682773533116438
MAE: 8.49694574723678
```

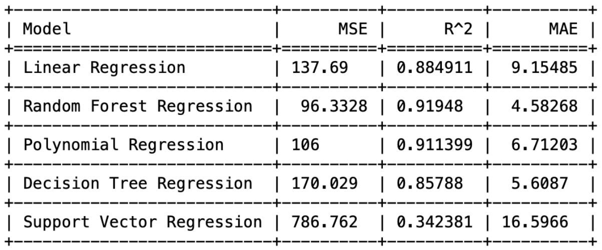
Fig. 2: Improved PR after feature selection



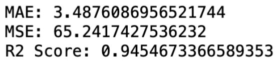
```
MAE: 9.54911662593501
MSE: 170.75138988946864
R2 Score: 0.8227144378755629
```

Fig. 3: Worsened RFR after feature selection

7) Time Series Analysis on Original Data:

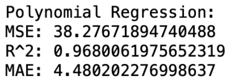


```
+---------------------------+----------+----------+----------+
| Model                     |      MSE |      R^2 |      MAE |
+===========================+==========+==========+==========+
| Linear Regression         | 137.69   | 0.884911 |  9.15485 |
+---------------------------+----------+----------+----------+
| Random Forest Regression  |  96.3328 | 0.91948  |  4.58268 |
+---------------------------+----------+----------+----------+
| Polynomial Regression     | 106      | 0.911399 |  6.71203 |
+---------------------------+----------+----------+----------+
| Decision Tree Regression  | 170.029  | 0.85788  |  5.6087  |
+---------------------------+----------+----------+----------+
| Support Vector Regression | 786.762  | 0.342381 | 16.5966  |
+---------------------------+----------+----------+----------+
```



```
MAE: 3.4876086956521744
MSE: 65.2417427536232
R2 Score: 0.9454673366589353
```

Fig. 4: Improved RFR after feature selection

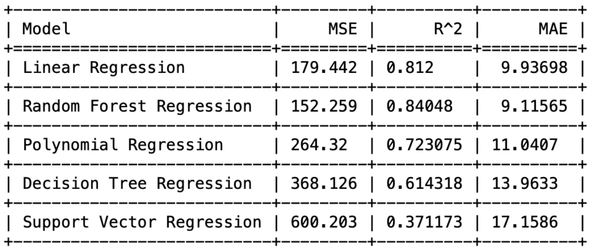


```
Polynomial Regression:
MSE: 38.27671894740488
R^2: 0.9680061975652319
MAE: 4.480202276998637
```

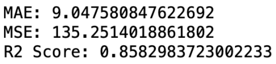
Fig. 5: Improved PR after feature selection

It is clear that Polynomial Regression performed the best when it came to time series analysis on Original Data. Use of Feature selection further improved the results.

8) Time Series Analysis on Residual Data:

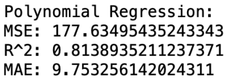


```
+---------------------------+---------+----------+----------+
| Model                     |     MSE |      R^2 |      MAE |
+===========================+=========+==========+==========+
| Linear Regression         | 179.442 | 0.812    |  9.93698 |
+---------------------------+---------+----------+----------+
| Random Forest Regression  | 152.259 | 0.84048  |  9.11565 |
+---------------------------+---------+----------+----------+
| Polynomial Regression     | 264.32  | 0.723075 | 11.0407  |
+---------------------------+---------+----------+----------+
| Decision Tree Regression  | 368.126 | 0.614318 | 13.9633  |
+---------------------------+---------+----------+----------+
| Support Vector Regression | 600.203 | 0.371173 | 17.1586  |
+---------------------------+---------+----------+----------+
```



```
MAE: 9.047580847622692
MSE: 135.2514018861802
R2 Score: 0.8582983723002233
```

Fig. 6: Improved RFR after feature selection



```
Polynomial Regression:
MSE: 177.63495435243343
R^2: 0.8138935211237371
MAE: 9.753256142024311
```

Fig. 7: Improved PR after feature selection

It is clear that Random Forest Regression performed the best when it came to time series analysis on Residual Data. Use of Feature selection further improved the results.

9) Analysis: In the classification tasks, the three top performing models were seen to be Decision Tree, Random Forest and Gradient Boost. Random Forest and Gradient Boost are essentially an ensemble of multiple Decision trees and hence the fact that all three of them produce good results shows the effectiveness of Decision Tree in this particular task. This can be attributed to the ability of Decision Trees to capture non linear relationship between variables and to carry an inherent form of feature selection as well. This inherent feature selection also meant that te external feature selection carried out was not too effective, although that is also attributed to the low dimensionality of the data. The application of SMOTE as part of our analysis successfully addressed the class imbalance issue by oversampling the minority class. This technique provided a more balanced representation of the classes, leading to improved classification results and enabling the classifier to make better predictions for high pollution levels despite the initially imbalanced dataset. For regression, Polynomial Regression performed the best overall, suggesting that there are non-linear relationships between the features and pollution levels in the data. Polynomial Regression's flexibility in capturing higher-order interactions allows it to model these complex patterns effectively. On the other hand, Random Forest Regression showed better performance in the temporal analysis, indicating its ability to capture temporal dependencies and handle high-dimensional data. Time series data often exhibit temporal dependencies, where the value at one time point is related to previous values. Random Forest Regression may have been more successful in capturing these temporal patterns due to its ability to consider the historical context when making predictions. The ensemble nature of Random Forest Regression also contributes to its effectiveness. By combining multiple decision trees, it can reduce overfitting and improve generalization.

V. Conclusion

After much testing and application of models, it was found that Polynomial Regression based on Original data performs the best when it comes to regression, and Gradient Boost applied in tandem with SMOTE and feature selection gives the highest classification accuracy. There were a number of observtions made about the relationship between variables, such as the fact that O3 is heavily correlated with AQI level when the level of O3 is abobe a certain threshold. There were also instances of peculiar correlation being verified by a general search, such as the fact that NO2 is negatively correlated with wind speed. The best models were able to achieve an accuracy of 0.986532994164298 when it came to classification, a the lowest MAE value of 9.04 when it came to time series analysis. Some other paths for furthering the cause would be to take into account the effects of Public Holidays, Rainy Days, Fuel Prices etc. This will enable the administration to make policies that could actually make a difference.

References

[1] Shwet Ketu and Pramod Kumar Mishra. "Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare". In: *Complex & Intelligent Systems* 7.5, 2597 (2021), pp. 2597–2615.

ISSN: 2198-6053. DOI: 10.1007/s40747-021-00435-5. URL: https://doi.org/10.1007/s40747-021-00435-5.

[2] Zixi Zhao et al. “A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic”. In: *Scientific Reports* 13.1, 1015 (2023), p. 1015. ISSN: 2045-2322. DOI: 10.1038/s41598-023-28287-8. URL: https://doi.org/10.1038/s41598-023-28287-8.