Comparing Various AI/ML Techniques to Predict Air Pollution

Sarvesh Dalvi[1], Satyajit Sahu[1], Abhishek Mishra[1], Vaishali Gatty[2]

[1]Student, Department of MCA, Vivekanand Education Society's Institute of Technology, Maharashtra, India.
[2]Assistant Professor, Department of MCA, Vivekanand Education Society's Institute of Technology, Maharashtra, India.

---***---

Abstract - *Air pollution in general is a problem on a global scale.Developing countries like India are the ones which are most impacted by the pollution. India ranks 3rd out of 106 countries in terms of air pollution. Air Quality Assessment can be done with the help of Air Quality Index Parameters which can be used for assessing the air pollution hot spots in the region and taking necessary actions to curb it.This paper discusses the rising air pollution in the country and the negative effects that are associated with it.*
It also includes various models which predict the trends in air pollution in the city using various pollutant concentrations and AQI as a parameter.

***Key Words*: Artificial intelligence, Machine Learning, Air Pollution, Prediction, Graphs, Plots.**

1.INTRODUCTION

Air pollution refers to the release of pollutants into the air—pollutants which are detrimental to human health and the planet as a whole.Air pollution is responsible for nearly seven million deaths around the world yearly, according to the World Health Organization (WHO). 9 off 10 human beings currently breathe air that exceeds the WHO's guideline limits for pollutants and the people who are living in low and middle income countries suffering the most effects resulting in adverse effect on the public's health.[1]

Air pollution poses a great threat to the health of the people as they are exposed to fine particles in polluted air that go into the lungs and cardiovascular system, causing numerous diseases like stroke, COPD, various heart diseases and complications, lung cancer, and respiratory ailments/infections.Industries, all modes of transportation, coal and other fossil fuel power plants and household solid fuel usage are by far one of the major contributors to air pollution. Air pollution in the country continues to rise at an unprecedented rate and is rapidly affecting economies and people's quality of life.[2]

Artificial intelligence, in simple terms, is essentially a simulation of human intelligent processes reproduced by machines, namely computer systems.Artificial Intelligence is usually used for natural language processing(NLP), speech recognition and machine vision.

AI systems function by accepting large amounts of training data which is then analyzed for correlations and patterns, and then using these patterns to make predictions about future variables. To give an example, a programmed chatbot that is fed examples of text chats can learn to communicate with people with indistinguishable temperament of humans, or an image recognition tool can could learn to identify and describe objects in images by reviewing millions of sample images given to it.[3]

Machine learning, in simple terms, is a branch of computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Supervised learning is also known to be supervised machine learning, and is defined by its use of labeled datasets which are used to train algorithms that classify data or predict outcomes with good accuracy.When the input data is fed into the model, it adjusts its weight until the model has been fitted properly. This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting. Supervised learning helps various organizations solve a variety of problems in the real-world, such as creating a separate folder for spam from your inbox. Some methods used in supervised learning include naïve bayes, support vector machine (SVM),logistic regression, linear regression,random forest,neural networks, and more.[4]

The model implemented here is easy to use and predicts the air pollution status of a city(here, Mumbai) with good accuracy. It considers the historical data of air pollution and predicts results based on air quality index. The main objective is to provide the people with information on the air quality. No organization can claim for quick response and easily solve the issue of air pollution. It's an established fact that air pollution affects a large number of people and cannot be stopped suddenly one day.In order to make the air quality better a lot of small baby steps need to be taken. The application provides us with a critical view of the current state of the cities in India

It is ensuring that more and more people are coming forward so as to secure the environment and familiarize all with the harmful effects brought by air pollution.

2. LITERATURE SURVEY

Recently India has been making efforts to reduce air pollution.India Recently signed COP21 Accord in order to curb their carbon emissions by 2025.

In a published paper[5] it was concluded that, "....for cities like Pune ,Mumbai where concentration of so2 is increasing we can take measures from now to not face problems later."

Despite taking various measures India has been seeing a steady incline in the air pollution levels.Reports show that the levels of air pollutants are on the rise in most major cities including the National Capital of the country,New Delhi which was ranked as most polluted city in the world by WHO in 2018.The estimated losses caused by air pollution is around 5-7 lakh crore annually.

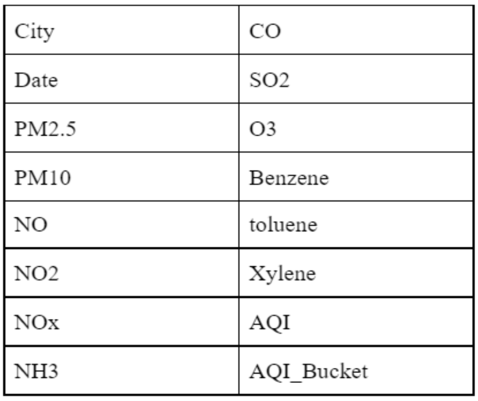
51% of pollution is caused by industrial pollution, 27 % by vehicles, 17% by crop burning and 5% by fireworks. Air pollution contributes to the premature deaths of 2 million Indians every year.

At the beginning of 2019 the Indian government inaugurated the National Clean Air Program (NCAP) to address the situation. It is their aim to reduce levels of air pollution by 20-30 per cent by 2024 in over 122 of the worst affected cities. Actions being taken in New Delhi, Ahmedabad and Pune include the implementation of health risk communications plans, the increase in the number of monitoring stations and better control of industrial emissions.[6]

Though it was observed in a recently published paper[7] that "the air pollutant concentration has reduced in every city of the world during the lockdown period. It has been also detected that the PM 2.5 and PM 10 are the most affecting air concentrator which controls the air quality of all the selected places during and after lockdown."

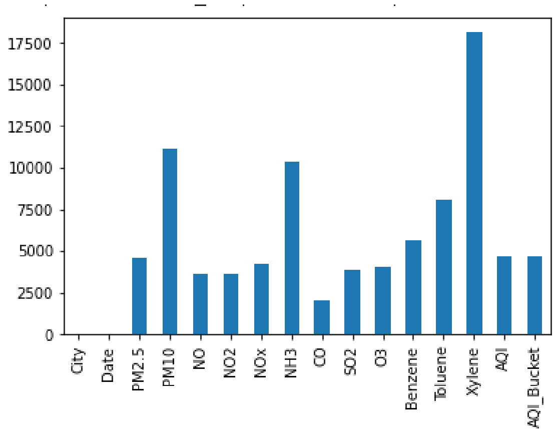
3. DATASET

Source: Kaggle
Structured / Unstructured Data: The data is structured in csv format.
Dataset: The dataset contains ~20000 records of major cities of India.The dataset used here contains the following attributes listed below:



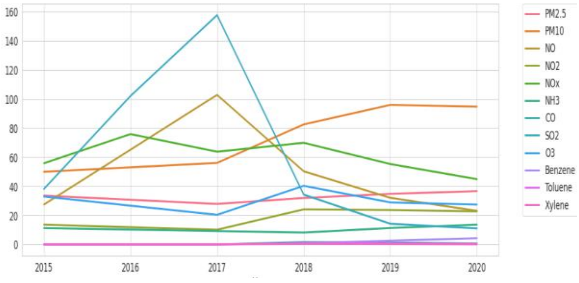
City	CO
Date	SO2
PM2.5	O3
PM10	Benzene
NO	toluene
NO2	Xylene
NOx	AQI
NH3	AQI_Bucket

Preprocessing and cleaning of data: The preprocessing and cleaning of data was done before performing exploratory data analysis. The missing data from each of the dataset was mapped and filled in. About 25% of missing data was filled.

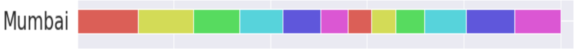


4. EXPLORATORY DATA ANALYSIS

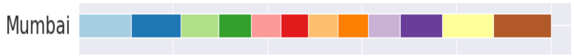
The results for exploratory data analysis for three cities were done. The reports for Mumbai are represented by the following results.



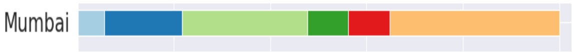
CO level across months in Mumbai



PM 2.5 levels in Mumbai 2015-2020(Average)



AQI bucket for Mumbai 2015- 2020(Average)



Each color in the above figures represent the 12 months sequentially.

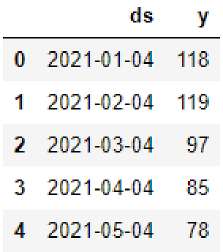
5. PROPOSED MODELS AND THEIR RESULTS

5.1 PROPHET PROCEDURE

Prophet is a model for forecasting time series data based on an additive model where non linear data can be fit into monthly,weekly and daily trends.
Prophet is an open source software which was developed and released by Facebooks's own Core Data Science team. The effectiveness of this model is best if the series has ample amount of historical data.
This is what makes Prophet Procedure a great candidate model to predict air pollution based on date and historical values.[8]
The following are some of the key advantages of This model:

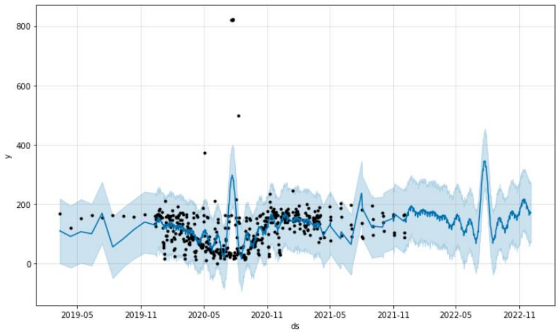
- **Accurate and fast**
- **Fully automatic**
- **Tunable forecasts**
- **Available in python and R**

prophet model is applied to the AQI dataset* for Mumbai.

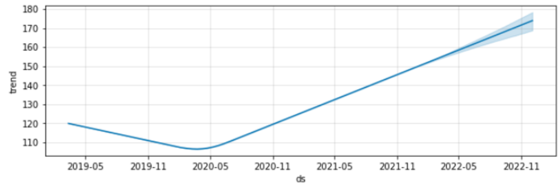


	ds	y
0	2021-01-04	118
1	2021-02-04	119
2	2021-03-04	97
3	2021-04-04	85
4	2021-05-04	78

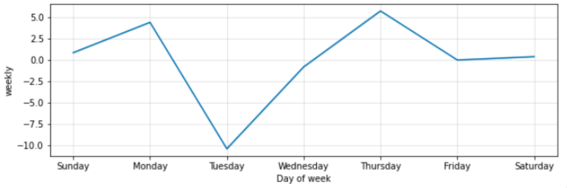
This dataset contains ~800 entries Ds is the date value and y is AQI value The columns had to be renamed because the model only accepts 'ds' and 'y' as names. The following is the outcome is achieved with this model:



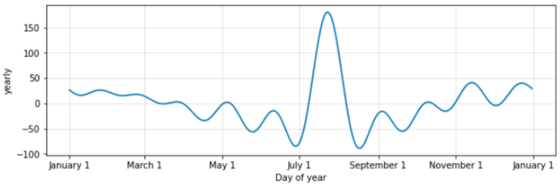
The model handled the outliers quite effectively without deviating from the trend too much.The following are the various trends are obtained from the model:



The above graph shows the trend over the years.



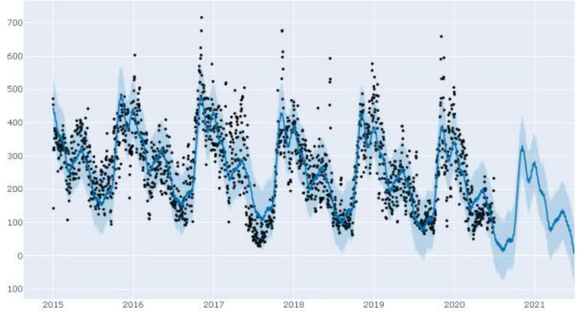
The above graph shows weekly changes in the AQI.

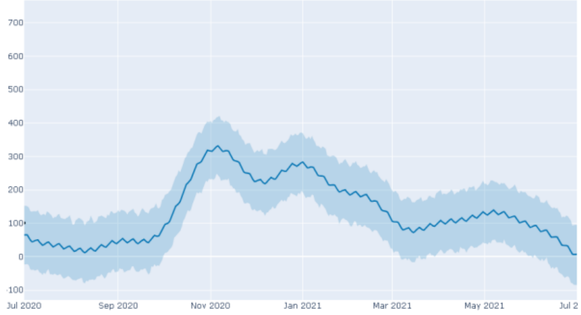


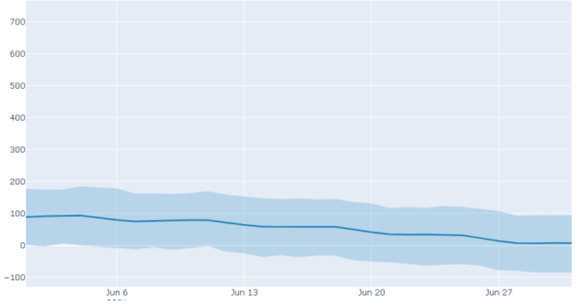
In conclusion, the Prophet procedure is Effective at predicting trends and shows detailed statistics which is quite useful in predicting air pollution.
The only drawback is that only one city at a time can be processed at a time using this model.

5.2 PROPHET ANALYSIS USING DIFFERENT DATASET

The following results are obtained by using a different dataset. The result shows yearly and monthly predictions.







5.3 ARIMA ANALYSIS

ARIMA is short form for Autoregressive Integrated Moving Average.

ARIMA is a statistical analysis model which uses time series data to either better understand the dataset or predict future trends.

ARIMA is a type of regression analysis that compares one variable to another variable.

Arima works by predicting the data using the difference between two distinct data rather than using the actual values.[9]

A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt':

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

where, Y_{t-1} is the lag1 of the series, beta1 is the coefficient of lag1 that the model estimates and alpha is the intercept term, also estimated by the model.[10]

A pure Moving Average (MA only) model is one where Y_t depends only on the lagged forecast errors:

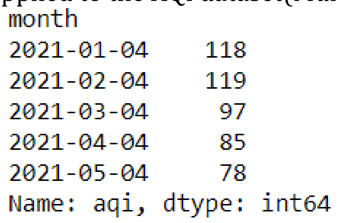
$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

An ARIMA model is one where the time series was different at least once to make it stationary and you combine the AR and the MA terms given above.The equation merges into:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

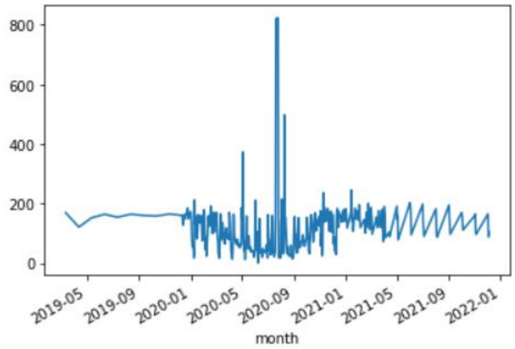
Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lag forecast errors.

The model is applied to the AQI dataset(Mumbai):



```
month
2021-01-04    118
2021-02-04    119
2021-03-04     97
2021-04-04     85
2021-05-04     78
Name: aqi, dtype: int64
```

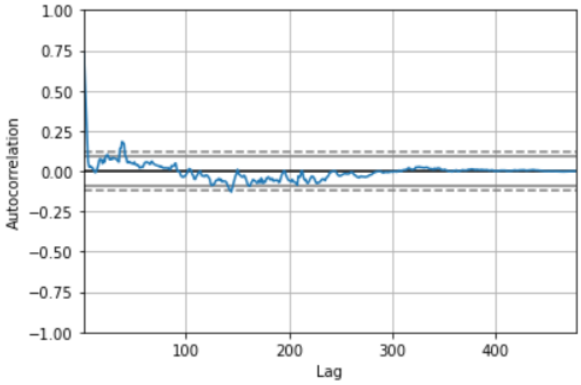
The data plot is as follows:



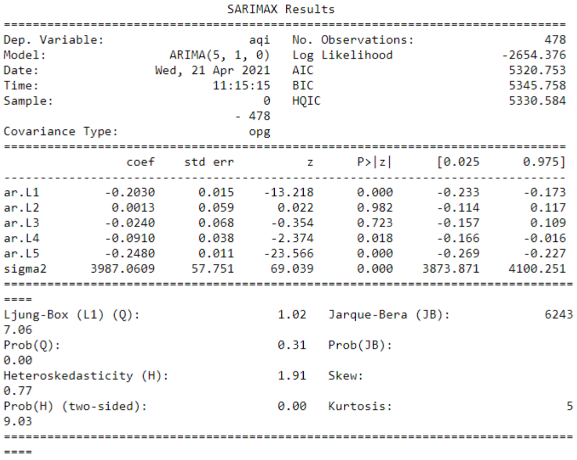
Using the model it is found that Autocorrelation aka serial correlation which is the correlation of the data with a delayed copy of itself as a delayed function.

Essentially to gauge the similarity between data as a function of the time lag between the two.

The graph describing the same is given below:

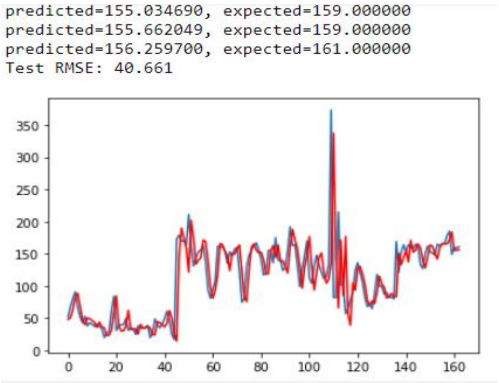


Seasonal Auto-Regressive Integrated Moving Average model details are given below:



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                    aqi   No. Observations:                  478
Model:                 ARIMA(5, 1, 0)   Log Likelihood               -2654.376
Date:                Wed, 21 Apr 2021   AIC                           5320.753
Time:                        11:15:15   BIC                           5345.758
Sample:                             0   HQIC                          5330.584
                                - 478
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.2030      0.015    -13.218      0.000      -0.233      -0.173
ar.L2          0.0013      0.059      0.022      0.982      -0.114       0.117
ar.L3         -0.0240      0.068     -0.354      0.723      -0.157       0.109
ar.L4         -0.0910      0.038     -2.374      0.018      -0.166      -0.016
ar.L5         -0.2480      0.011    -23.566      0.000      -0.269      -0.227
sigma2      3987.0609     57.751     69.039      0.000    3873.871    4100.251
================================================================================
====
Ljung-Box (L1) (Q):                   1.02   Jarque-Bera (JB):              6243
7.06
Prob(Q):                              0.31   Prob(JB):
0.00
Heteroskedasticity (H):               1.91   Skew:
0.77
Prob(H) (two-sided):                  0.00   Kurtosis:                         5
9.03
================================================================================
====
```

The final prediction is as follows:



```
predicted=155.034690, expected=159.000000
predicted=155.662049, expected=159.000000
predicted=156.259700, expected=161.000000
Test RMSE: 40.661
```

RMSE stands for Root Mean Square Error.

Mean Square is an error metric for determining the accuracy and the error rate of the model.

RMSE is the Square Root of the value from Mean Square Error.

An RMSE score of less than 180 is considered a good score for a moderately or well fit algorithm. If the RMSE value exceeds 180,feature selection is necessary and hyper parameter tuning on the parameters of the mode is recommended.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:22.094

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Mean Squared Error (MSE) is the mean of the squared errors:1653.334

$$\mathrm{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

The RMSE value that is obtained is 40.661 which is well within the acceptable parameters.

5.4 LINEAR REGRESSION

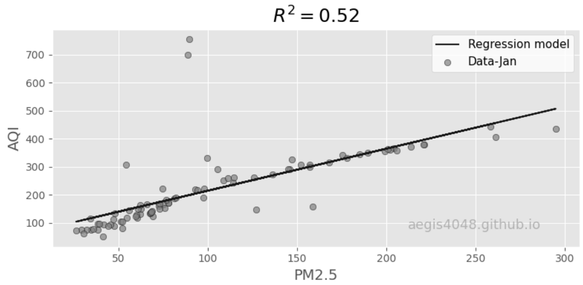
Linear regression is one of the most commonly known machine learning algorithms used for supervised learning. It is used to show and quantify the relationship between one or more variables which are used as predictors for the prediction of the subject (here air quality index).

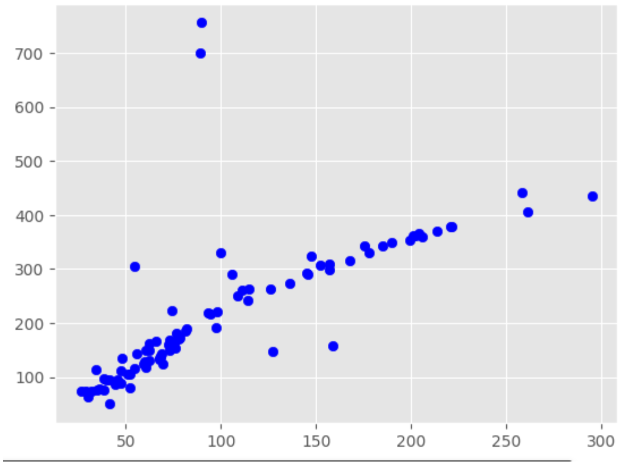
It is assumed that two variables are always linearly related. Hence, linear function that predicts the response value(y) as precisely and accurately as possible as a function of the feature or independent variable(X).[11]

The formula for simple linear regression is as follows:

$y = b_0 + b_1 * x_1$

The following figure shows the plot for AQI vs PM 2.5





There are three common evaluation metrics for linear regression problems:

Mean Absolute Error (MAE) is the mean of the absolute value of the errors: 115.53.

Mean Squared Error (MSE) is the mean of the squared errors:22448.323.

Root Mean Squared Error (RMSE) is the square root of the mean of the squared error:149.827

6. FINAL RESULTS OBTAINED IN THE 2 MODELS

	MAE	MSE	RMSE
ARIMA Analysis	22.094	1653.334	40.661
Linear Regression	115.53	22448.323	149.827

7. CONCLUSIONS

Based on the above models it is seen that a steady increase in pollutants based on AQI levels and that Mumbai is headed towards unprecedented levels of air pollution like the capital of India(Delhi) in near future. Development of various prediction models is crucial in understanding the current state of air condition in the city, studying various sources of pollutants and developing effective measures to curb them.This data can be used to spread further awareness about current pollution trends in the city.

The above models are based on the AQI and predict for a single city(Mumbai). The models show that the air quality is deteriorating which can give rise to various cardiovascular effects such as heart attacks, and respiratory problems like asthma.

Further research can be done to make the models more efficient and accurate for multiple cities.

8. REFERENCES

[1] Air Pollution in the western Pacific https://www.who.int/westernpacific/health-topics/air-pollution

[2] Air Pollution Facts, Causes and the Effects of Pollutants in the Air. https://www.nrdc.org/stories/air-pollution-everything-you-need-know

[3] What is artificial intelligence (AI)?-AI definition and how it works. https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence

[4] What is Machine Learning?-India |IBM https://www.ibm.com/in-en/cloud/learn/machine-learning

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.Gope, Sadhan & Dawn, Subhojit & Das, Shreya. (2021). Effect of COVID-19 pandemic on air quality: a study based on Air Quality Index. Environmental Science and Pollution Research. 10.1007/s11356-021-14462-9.

[6] India Air Quality Index(AQI) and Air Pollution Information | AirVisual:: https://www.iqair.com/india

[7] Gope, Sadhan & Dawn, Subhojit & Das, Shreya. (2021). Effect of COVID-19 pandemic on air quality: a study based on Air Quality Index. Environmental Science and Pollution Research. 10.1007/s11356-021-14462-9.

[8] Prophet | Forecasting at scale https://facebook.github.io/prophet/

[9] How to Create an ARIMA Model for Time Series Forecasting in Python https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[10] Autoregressive Integrated Moving Average (ARIMA) https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

[11] sklearn.linear_model.Linear Regression https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html