Air Quality Index Analysis and Prediction

Chakkiyath Jaisal Sabu, MSc. Big Data Analytics

***Abstract*— Air is one of the most vital element of nature that sustains life on earth and the monitoring , prediction and analysis of Air quality can be considered as one the great milestones that humans have achieved. The advancements in Machine learning , Artificial intelligence and other computing technologies have paved way for a huge leap in the overall analysis of Air Quality Index (AQI). Environment sustainability has been a key issue over many decades and research has been conducted on multiple fronts over the last two decades. Regular monitoring and analysis has and will always help scientists in understanding the upcoming threats and work hand in hand with governments and several non governmental organizations in taking actions and implementing restrictions for a quality future with a good AQI. The simple rule of prevention is better than cure is the crux of AQI analysis, getting to know the threats beforehand will prepare us to deal with it better. In the project we will be taking into account the air quality data published by Central Pollution Control Board Of India. We will be taking the data from 2015-2020 for our analysis and will apply several algorithms to forecast the air quality of India in coming years and perform a analysis of the data to infer several facts. The project includes all the major stages of a Data Analytic Pipeline and the detailed exploratory analysis of the data-set.**

***Index Terms*— Deep Learning(DL), Machine Learning(ML), Extreme Gradient Boosting (XGBoost), Seasonal Auto Regressive Integrated Moving Averages, Auto Regressive Integrated Moving Averages, Pyspark, Pandas, Air Quality Index(AQI)**

=

I. INTRODUCTION

THe Air Quality Index (AQI) is a parameter for measuring the quality of air taking into consideration multiple elements that contribute to the index. The lower value of AQI indicates a better air with lower amounts of pollutants and the quality of air decreases with a higher AQI. There are several pollutants that are took into consideration while calculating the Air Quality Index. PM2.5 and PM10 are particulates that is also known as Suspended Particulates Matter. These suspended particles have two different categories that has a diameter of 10 micrometer or less as well as particles with diameter of 2.5 micrometer or less. There are other components like Nitrogen Dioxide (NO2), Nitric oxide (NO), Nitrogen Oxides (NOx), Ammonia (NH3) and Carbon Monoxide (CO) released by the emissions from vehicles. Pollutants like Benzene, Toluene, Xylene, Sulphur Dioxide (SO2) and Ozone (O3) released by Industries is also considered while calculating the index. In our project we will be categorizing these pollutants and the analysing their impact on the air quality of cities in India. [6] The Air Quality Index values are measured on a scale like in the figure below. The simple rule applies that lower the AQI the better the air. Air pollution is one of the major factors of rising health issues among the population across the world. While considering our project , developing countries like India with a relatively higher number of pollutant sources like vehicles and industries have to develop a better forecasting system for the air quality to keep a check on the possible threats in future which in turn will allow them to take necessary measures. The Code and Documentation to the project can be found at https://github.com/jaisalsabu/BDATechnicalReport

Under the supervision of Dr. Shagufta Henna, Letterkenny Institute of Technology, Letterkenny, CO. Donegal

AQI	AQI Classes	Health Impact	Suggestions
0~50	Excellent	The air quality is satisfactory	It is suitable for normal actions for various people.
51~100	Good	Have weak health effects on extremely sensitive people	Extremely sensitive people should reduce outdoor activities.
101~150	Light pollution	Healthy people show signs of irritation	Children, the elderly and patients with heart disease should reduce outdoor activities.
151~200	Moderate pollution	It may affect the heart and respiratory systems of healthy people	Even healthy people should reduce outdoor sports activities.
201~300	Serious pollution	The symptoms of heart disease and lung disease increased significantly	Children, the elderly and patients with heart disease should stop outdoor activities.
201~300	Heavy pollution	Healthy people have obvious strong symptoms	Healthy people should avoid outdoor activities.

Fig. 1. Air Quality Index Scale.

II. BACKGROUND

While taking into consideration the prediction models built, we can see a wide range of time series forecasting algorithms ranging from ARIMA to VARIMAX . Other Machine learning and deep learning algorithms like Adaboost, Catboost etc have also been used widely by programmers and scientists. Mauro Castelli in one of his recent studies compared the ARIMA model with Holt Exponential Smoothing [1]. In a paper published by Chen, a a model was predicted using XGBoost and the advantages of the model was clearly explained [4].In another model Proposed by Q Huang a model based on SVR is proposed to predict the Air quality in China and his team has successfully built a working efficient model [5].Lastly we also took into consideration the Double Exponential smoothing prediction model built by Bose [7].

III. DATA-SET DESCRIPTION

Air Quality Index dataset(2015-2020) sourced from Central Pollution Control board of India [2] [3] is the core of the technical project. The Dataset is a combination of 5 csv file named city_hour.csv , city_day.csv, station_day.csv, station.csv and station_hour.csv . Out of these we will be using three

```
root
 |-- City: string (nullable = true)
 |-- Date: timestamp (nullable = true)
 |-- PM2.5: double (nullable = true)
 |-- PM10: double (nullable = true)
 |-- NO: double (nullable = true)
 |-- NO2: double (nullable = true)
 |-- NOx: double (nullable = true)
 |-- NH3: double (nullable = true)
 |-- CO: double (nullable = true)
 |-- SO2: double (nullable = true)
 |-- O3: double (nullable = true)
 |-- Benzene: double (nullable = true)
 |-- Toluene: double (nullable = true)
 |-- Xylene: double (nullable = true)
 |-- AQI: double (nullable = true)
 |-- AQI_Bucket: string (nullable = true)
```

Fig. 2. Schema for Dataframe.

for our project. All the datasets sourced for the project is structured and has a well defined schema at with all the datatypes. PM10, PM2.5, NO2, SO2, CO, O3, NH3, Pb, AQI, City, Date and AQI Bucket are the columns in the main csv file. The cities took into consideration for calculating the AQI are Ahmedabad,Gurugram,Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal, Guwahati, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Kochi, Ernakulam,, Hyderabad, Talcher, Jaipur, Jorapokhar, Kolkata, Lucknow, Mumbai, Patna, Shillong, Thiruvananthapuram, Visakhapatnam . There are two categorical columns in the dataset and other columns are numerical and Date values. There are multiple rows with null values which needs to be imputed with mean or dropped from the dataframe before working with the analysis.The goal of technical project can be divided into two parts Analysis and Forecasting. In the analysis part we will be performing Exploratory data analysis on the dataset to find out the cities with highest pollution as well as identify the major pollutants that contribute to Air Quality Index Value. In the second part we will be developing a forecasting model with SARIMA (Seasonal Auto Regressive Integrated Moving Average) and XGBoost to effectively predict the Air Quality Index of India for the coming years. throughout the entire project we will be following a Data Analytic Pipeline with well defined stages.

IV. GOAL 1-ANALYSIS

A. Data Acquisition

The very first stage in a analytic pipeline is loading the data from csv file and forming a dataframe. since we are using the city day data we have to load the city_day.csv dataset. We read the csv file as a Spark Dataframe named df for further processing. We can have a glimpse of the first seven rows of the dataframe in Fig 4.

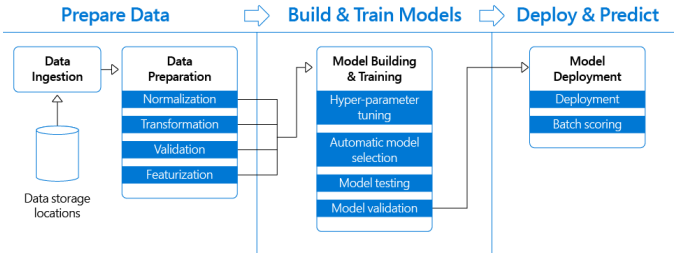

Fig. 3. Stages in Data Analytic Workflow.

	City	Date	PM2.5	PM10	NO	NO2	NOx
1	Ahmedabad	2015-01-01T00:00:00.000+0000	null	null	0.92	18.22	17.15
2	Ahmedabad	2015-01-02T00:00:00.000+0000	null	null	0.97	15.69	16.46
3	Ahmedabad	2015-01-03T00:00:00.000+0000	null	null	17.4	19.3	29.7
4	Ahmedabad	2015-01-04T00:00:00.000+0000	null	null	1.7	18.48	17.97
5	Ahmedabad	2015-01-05T00:00:00.000+0000	null	null	22.1	21.42	37.76
6	Ahmedabad	2015-01-06T00:00:00.000+0000	null	null	45.41	38.48	81.5
7	Ahmedabad	2015-01-07T00:00:00.000+0000	null	null	112.16	40.62	130.77

Fig. 4. First 7 rows of the df dataframe .

```
root
 |-- City: string (nullable = true)
 |-- Date: timestamp (nullable = true)
 |-- AQI: double (nullable = false)
 |-- AirQuality: string (nullable = true)
 |-- VehiclePollutants: double (nullable = false)
 |-- IndustrialPollutants: double (nullable = false)
```

Fig. 5. Schema of transformed dataframe .

B. Data Transformation

To perform an effective descriptive analysis we have prepare our data. we rename the columns AQI_Bucket and PM2.5 to AirQuality and PM25 respectively. We also delete the duplicate rows which can hinder the results from the dataframe. In the projects reviewed earlier as a part of background study for the technical project, many of the users have deleted the rows with null values which affects the accuracy of the results. So we will impute the null values with the mean of that column as a measure to improve the accuracy of analytics and the prediction. The other major stage in Exploratory analysis is the combining of the pollutants into two categories VehiclePollutants and IndustrialPollutants respectively. We have to merge the columns PM25,PM10,NO,NO2,NOx,NH3 and CO columnns and create the new column VehiclePollutants. The sum of the values of these columns will be added to create a value for the VehiclePollutants. In the same way we will also add the columns SO2,O3,Benzene,Toluene and Xylene to form the new column Industrial Pollutants. since we have created the two new columns we will drop the remaining columns that do not add to the value. The schema of transformed dataframe is provided in the Fig.5.

C. Descriptive Analysis

After pre-processing and transforming our data, we have to perform the descriptive analysis and prove two of our hypothesis in this stage. As for our first hypothesis of finding

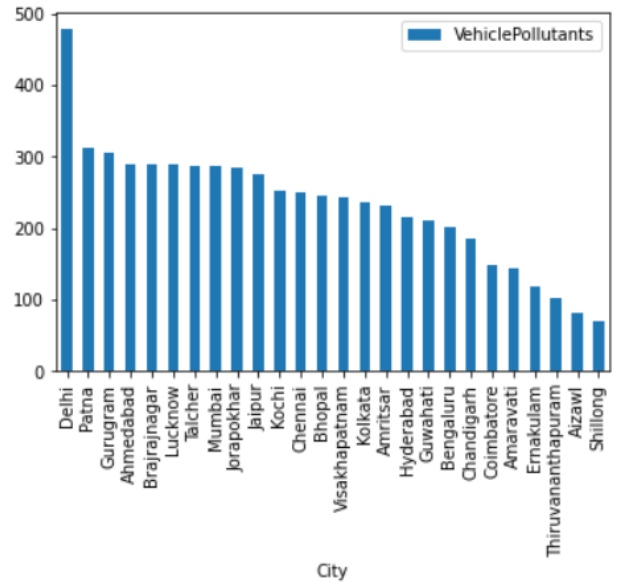

Fig. 6. Vehicle Pollution plot.

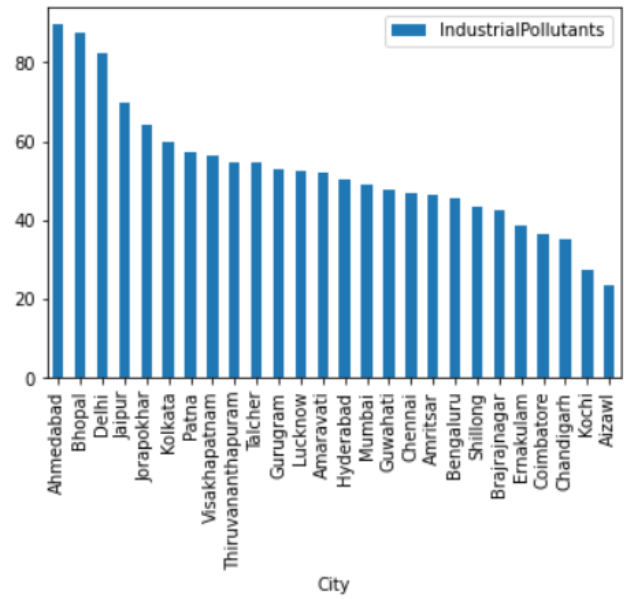

Fig. 7. Industrial Pollution plot.

out the most polluted cities for each category of pollutant. we will plot the graph of Vehicle Pollutants and Industrial Pollutants based on cities to find out the cities with highest Value. For better visualization techniques we will convert the spark dataframe to Pandas.

D. Results

1) CITY WITH HIGHEST VEHICLE POLLUTION:

a) From the Fig.6 we can deduce that Delhi is the city with highest Vehicle pollution followed by Patna and Gurugram.:

2) CITY WITH LOWEST VEHICLE POLLUTION:

a) From the Fig.6 we can infer that Shillong is the city with lowest Vehicle Pollution followed by Aizwal and Thiruvananthapuram.:

3) CITY WITH HIGHEST INDUSTRIAL POLLUTION:

a) From the Fig.6 we can deduce that Ahmedabad is the city with highest Industrial pollution followed by Bhopal and Gurugram.:

4) CITY WITH LOWEST INDUSTRIAL POLLUTION:

a) From the Fig.6 we can infer that Shillong is the city with lowest Vehicle Pollution followed by Aizwal and Thiruvananthapuram.:

V. GOAL 2- FORECASTING

A. SARIMA (SEASONAL AUTO REGRESSIVE INTEGRATED MOVING AVERAGES)

ARIMA (AUTO REGRESSIVE INTEGRATED MOVING AVERAGE) is one of the most widely used algorithms for forecasting. Its used in business demand forecasting, ustomer churn forecasting and many other sectors effectively. But while taking into consideration the the data with a seasonal component there are limitations to ARIMA model. Even though the ARIMA Model handles data with trend, it is not effective in handling data with a seasonal component. Seasonal component is a simple time series but with the inclusion of a repetitive cycle.

1) IMPLEMENTATION:

a) Data Acquisition and Preprocessing: As we did in the earlier stage of descriptive analysis, we have to import the csv file city_day.csv and create a spark data frame and we have to drop duplicate and impute the mean and median for the null row values. As a par of the transformation process we also change the datatype of Date column to datetime format in pandas using the to_datetime() function.In an attempt to produce montb wise data in our dataframe we centerpoint our city column. The pivot_table() function is used for this purpose. In the later path, we have to predict the Air quality Index of India in total, so we have to create anew column AQI_India and add the sum of AQI from each city to the total.

b) Visualization: To get a clearer picture of our data , we have to plot the AQI of India. We plot a line graph of the Air Quality Index using matplotlib. From this graph in Fig.8 we see a downward trend and we can also deduce a seasonality from the graph. a repetitive cycle is clearly visible in our dataset. To get a more clearer picture of the seasonality trend and residual, a decomposed graph has to be made. From Fig.9 we can analyse the trend seasonality and residuals to understand our data better.

c) Augmented Dickey Fuller Test: Dickey-Fuller test is a unit root test that tests the mull hypothesis alpha is the coefficient of the first lag on Y. Augmented Dickey fuller test is a augmented version of the dickey fuller test and we add a differencing term to the equation of dickey fuller test. We perform this test on our model to understand the stationarity of our the time series. After performing the ADF test we get the results as in Fig.10. in the test statistics we have a p-value of 0.94, from which we understand that our data is not stationary. Now as a step make our data stationary, we

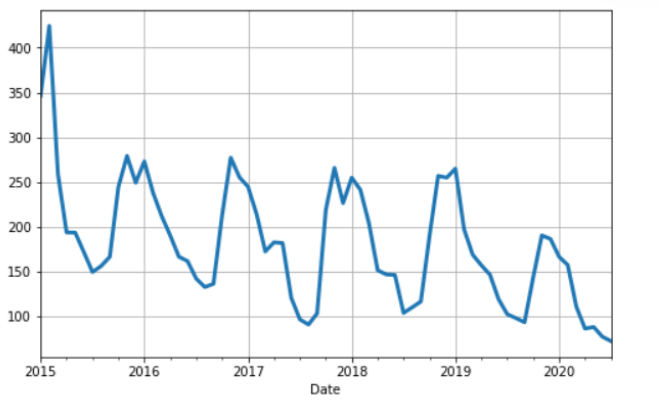

Fig. 8. Air Quality Index of India.

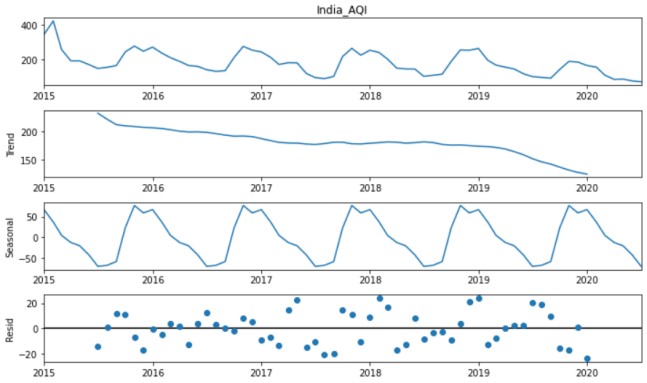

Fig. 9. Decomposed graph of trend, seasonality and residuals

introduce differencing of first order on our dataset and repeat the test. In the second test we get a value of 2.4 and we dedcue the value of d as 1

d) Implementing SARIMA: After we the desired results in the ADF test , we will use Auto-ARIMA to identify the best parameters for SARIMA model and the results are mentioned in Fig.11. from the results we infer that the model has an RMSE value of 22.75 and it can be used to forecast the values in our project

e) Forecasting: We split our datset into two sets at this stage into train and test sets, we train our model with the train set and the run our model on test set to get the results Using our model to forecast the values , we get the following results In Fig.12 we have efficiently predicted the Air Quality Index of India for 2022, and while observing the graph we find out that the year which experienced Covid lockdown is as major reason for our prediction being not accurate to the point. The closure of industries and reduced use of transport vehicles during the lockdown period has greatly influenced the Air Quality Index of India.

B. XGBOOST REGRESSOR

ABOUT XGBOOST is a implementation of the gradient boosting algorithm that turns out to be pretty effective. In a very short span of time it was popular among scientists and

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	67
Model:	SARIMAX(0, 1, 2)x(1, 0, [1], 12)	**Log Likelihood**	-316.908
Date:	Tue, 15 Feb 2022	**AIC**	643.816
Time:	15:53:48	**BIC**	654.765
Sample:	0	**HQIC**	648.143
	- 67		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.0189	0.059	0.320	0.749	-0.097	0.135
ma.L2	-0.8363	0.069	-12.077	0.000	-0.972	-0.701
ar.S.L12	0.9444	0.062	15.221	0.000	0.823	1.066
ma.S.L12	-0.5623	0.229	-2.458	0.014	-1.011	-0.114
sigma2	694.3700	142.982	4.856	0.000	414.130	974.610

Ljung-Box (L1) (Q):	0.95	**Jarque-Bera (JB):**	2.99
Prob(Q):	0.33	**Prob(JB):**	0.22
Heteroskedasticity (H):	0.38	**Skew:**	-0.52
Prob(H) (two-sided):	0.03	**Kurtosis:**	2.99

Fig. 10. SARIMAX Model Results

students for a wide range of applications. XGBoost is mainly used for regression predictive modelling. XGBoost is also known as Extreme Gradient Boosting and can be implemented using the scikit-learn API.

1) IMPLEMENTATION:

a) Data Acquisition: The XGBOOST model will be implemented in spark dataframes, so as we will load the city_hour.csv file and create a spark dataframe.The dataset has 707875 rows in total.

b) Data preprocessing: In the pre-processing stage we have to rename the columns AQI_Bucket and PM2.5 to AirQuality and PM25. As another step in transformation we will drop all the duplicate values and the impute the mean values like we did it for the Exploratory Data Analytics. Now we have to split the date column into 4 parts year, month, day and time for which we use the sql split function. Now we have to drop the AQI, Time and Datetime columns and we also change the datatype of Year, Month and Day column using cast function.

c) Implementing the model: First we split our dataset to train and test on a 7:3 ratio that will provide us with 495450 rows in training set and 212425 in testing set. To get clearer picture , we will plot the training data with City and AQI as values. We use a StringIndexer, VectorIndexer and Vector Assembler as a part of feature engineering. Now, We have to import the XGBoost regressor from sparkdl.xgboost and we create an instance of it and substitute AQI as the label column . While doing this we are creating the model training stage of pipeline. The model will take in the features of input column and will learn to predict the AQI

d) Tuning and Evaluation: We define a grid of hyper parameters to maxDepth and maxIter and an crossvalidator evaluation metric that compares the true labels with predicted values. This crossvalidator tunes the model.

e) Creating Pipeline: we create a pipeline with the stages defined as in the earlier segments. after we set the stages in pipeline we train the model with fit() command. The pipeline executes feature engineering, model tuning and training and then provides us with the best model as result

f) Prediction: In the last step of the model building we will make the needed predictions and run our model on the test

```
Out[93]: Test Statistic                  -0.114224
p-value                          0.948003
#Lags Used                      10.000000
Number of Observations Used     56.000000
Critical Value (1%)             -3.552928
Critical Value (5%)             -2.914731
Critical Value (10%)            -2.595137
dtype: float64
```

Fig. 11. ADF test results before differencing.

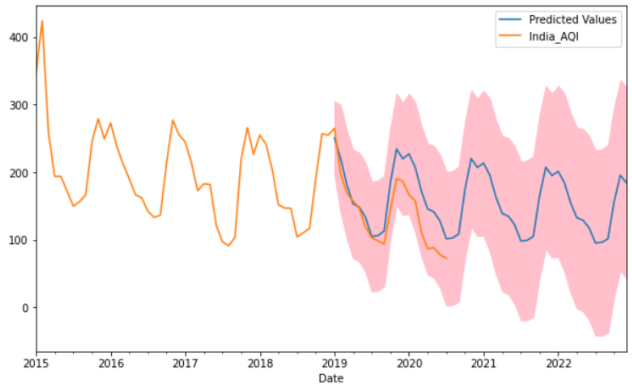

Fig. 12. AQI of India till 2022 forecasted with SARIMA.

data and evaluate it. After running the prediction we get the results and we also have observed an RMSE score of 76 and the Fig.13 shows the plotting of the predicted values against years.

VI. Results and Analysis

The quintessential aim of our technical project to use Time series algorithms in the forecasting of Air Quality Index has lead us to the implementation of two models SARIMA which expands to Seasonal Auto Regressive Integrated Moving Averages and the Extreme Gradient Boosting(XGBoost). We have implemented both the models using two different technologies pandas and pyspark. The SARIMA model that is actually an extension to the prevalent and widely used ARIMA model has a better RMSE score than the XGBoost model that was implemented on pyspark dataframes. While shortlisting the models to be used for the prediction the main factor that was to be considered is the seasonal component in the data . ARIMA (Auto Regressive Integrated Moving Averages) hand the very limitation of not being able to handle the seasonal component in the data. The other models that were considered were LSTM(Long Short Term Memory), LightGBM and VARIMA models. We have used SARIMA model to predict the Air Quality Index of India in 2021 and 2022. The Year 2020 had undergone a unexpected shift and many cities had a better Air Quality owing to the lockdown implemented in the wake of Covid-19. From Exploratory Data Analysis performed we have found out the cities with higher and lower levels in Industrial as well as Vehicle pollution contents and the results are noted in the respective section of the report. The Air Quality Index of India that we have predicted using the SARIMA model clearly portrayed in Fig.12 and we can note a slightly negative trend in the AQI of the country in total. However, due to the unexpected fluctuation of AQI values , the assurance of the an exact forecast cannot be guaranteed. The XGBoost model was also tuned using cross validator and an effective pipeline was built. One change noted in the the RMSE of XGBoost is that it increases with each run and can be fine-tuned to generate a better prediction model.

VII. Conclusion

The models we have implemented in the technical project can be improved consistently by combing two or more models to create a new model that can forecast the AQI with much better accuracy. Throughout the entire project we have implemented multiple technologies prevalent in the Data Analytic industry and the project is by no means a complete review or analysis of all the models being used for forecasting. From the technical project, we have successfully built a model that will predict the Air Quality Index of India as the dataset is sourced by the Central Pollution control Board in India. This model can be rebuiit to satisfy the requirements of different data sources and then generate a effective model.

References

[1] Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, Leonardo Vanneschi, "A Machine Learning Approach to Predict Air Quality in California", Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020. https://doi.org/10.1155/2020/8049504
[2] https://cpcb.nic.in/
[3] https://www.kaggle.com/rohanrao/air-quality-data-in-india
[4] Chen, Tianqi and Guestrin, Carlos.KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data MiningAugust 2016 Pages 785–794
[5] Q. Huang, J. Mao, and Y. Liu, "An improved grid search algorithm of SVR parameters optimization," in Proceedings of the International Conference on Communication Technology Proceedings, ICCT, pp. 1022–1026, Chengdu, China, November 2012.
[6] Nigam, S., R. B. K. N. and Mhaisalkar, V. (2015). Air quality index-a comparative study for assessing the status of air quality., Research Journal of Engineering and Technology pp. 267–274
[7] Bose, R., D. R. R. S. and Sarddar (2020). Time series forecasting using double exponential smoothing for predicting the major ambient air pollutants, In Information and communication technology for sustainable development,Springer,Singapore pp. 603–613.