

---

## **Time series forecasting of Air Quality Prediction**

Enrollment No.(s) - 19103175,19103070,19103071

Name of student(s) - Khushi Jain, Megha Agarwal, Pavini Jain

Supervisor name – Dr. Parul Agarwal

Submitted in evaluation of

MINOR PROJECT



in

Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

---

---

## Acknowledgement

It gives us immense pleasure to express our deepest sense of gratitude and sincere thanks to our highest respected and esteemed guide *Dr. Parul Agarwal* for her valuable guidance, encouragement and help for completing this work. Her useful suggestions for this whole work and co-operative behaviour are sincerely acknowledged. We would like to express our sincere thanks to her for giving us this opportunity to undertake this project. We would also like to express our indebtedness to our parents as well as our family members whose blessings and support always helped us to face the challenges ahead.

---

## Table of Contents

Contents	Page No.
<i>List of Figures</i>	6
<i>List of Tables</i>	7
<i>Self-Declaration Form</i>	8-9
<i>1. Motivation behind making the project</i>	10
<i>2. Introduction</i>	11-12
<i>3. Problem Statement</i>	12
<i>4. Methodology</i>	
<i>4.1. Flowchart</i>	13
<i>4.2. Data Collection</i>	14
<i>4.3. Data Cleansing</i>	14-15
<i>4.4. Data Visualization</i>	16-18
<i>4.5. Data Preprocessing</i>	19
<i>4.6. Training Dataset Formation</i>	19
<i>4.7. Architecture</i>	
<i>4.7.1. Multiple Linear Regression (MLR)</i>	19
<i>4.7.2 Support Vector Regressor (SVR)</i>	19

---

4.7.3 Random Forest Algorithm	20
4.7.4. Gradient Boosting Algorithm	20
4.7.5. Recurrent Neural Network (RNN)	20
4.7.6. Long Short Term Memory (LSTM)	21
4.7.7. Final Proposed Architecture	21-23

## 5. Results

### 5.1. Predictions

5.1.1. Prediction by Multiple Linear Regression	23-24
5.1.2. Prediction by Support Vector Regression	24-25
5.1.3. Prediction by Random Forest	25-26
5.1.4. Prediction by Gradient Boosting	26-27
5.1.5. Prediction by LSTM MODEL 1	27-28
5.1.6. Prediction by LSTM MODEL 2	28-29
5.1.7. Prediction by LSTM MODEL 3	29-30
5.1.8. Prediction by LSTM MODEL 4	30-31

### 5.2. Errors and Losses

5.2.1 R2 Squared	31-32
5.2.2. Mean Absolute Error	32-33
5.2.3. Root Mean Squared Error	33-34

### 5.3. Comparisons

---

<i>5.3.1. Comparison between various Machine Learning Models</i>	<i>35</i>
<i>5.3.2. Comparison between various LSTM Models</i>	<i>35</i>
<i>6. Conclusion</i>	<i>36</i>
<i>7. References</i>	<i>37</i>

---

## List of Figures

Figure No.	Figure Title	Page No.
Figure-4.1	Overall design of project	13
Figure-4.2	Visualization of data before and after cleansing	15
Figure-4.3	Visualization of Daily Emission of pollutant	16
Figure-4.4	Visualization of Monthly Emission of pollutant	17
Figure-4.5	Visualization of Maximum Emission of pollutant	18
Figure-4.6	Structure of RNN	20
Figure-4.7	Structure of LSTM cell	21
Figure-4.8	Design of LSTM Model 1	22
Figure-4.9	Design of LSTM Model 2	22
Figure-4.10	Design of LSTM Model 3	22
Figure-4.11	Design of LSTM Model 4	23
Figure-5.1	Prediction by Multiple Linear Regression	23
Figure-5.2	Prediction by Support Vector Regression	24
Figure-5.3	Prediction by Random Forest	25
Figure-5.4	Prediction by Gradient Boosting	26
Figure-5.5	Prediction by LSTM MODEL 1	27
Figure-5.6	Prediction by LSTM MODEL 2	28
Figure-5.7	Prediction by LSTM MODEL 3	29
Figure-5.8	Prediction by LSTM MODEL 4	30
Figure-5.9	Comparison between R2 value of various ML Models	32
Figure-5.10	Comparison between R2 value of various LSTM Models	32
Figure-5.11	Comparison between MAE value of various ML Model	33
Figure-5.12	Comparison between MAE value of various LSTM Model	33
Figure-5.13	Comparison between RMSE value of various ML Model	34
Figure-5.14	Comparison between RMSE value of various LSTM Model	34

---

## List of Tables

Table No.	Table Title	Page No.
Table-5.1	Prediction by Multiple Linear Regression	24
Table-5.2	Prediction by Support Vector Regression	25
Table-5.3	Prediction by Random Forest	26
Table-5.4	Prediction by Gradient Boosting	27
Table-5.5	Prediction by LSTM MODEL 1	28
Table-5.6	Prediction by LSTM MODEL 2	29
Table-5.7	Prediction by LSTM MODEL 3	30
Table-5.8	Prediction by LSTM MODEL 4	31
Table-5.9	Comparison between various ML Models	35
Table-5.10	Comparison between various LSTM Models	35

---

## Student's Self Declaration for Open Source libraries and other source code usage in Minor Project



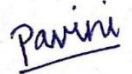
I/ We **Khushi Jain, Megha Agarwal, Pavini Jain(Group No. G64)** hereby declare the following usage of the open source code and prebuilt libraries in our minor project in **5<sup>th</sup> Semester** with the consent of our supervisor. We also measure the similarity percentage of pre written source code and our source code and the same is mentioned below. This measurement is true with best of our knowledge and abilities.

1. List of pre build libraries:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Pytorch
- Sklearn

2. List of pre build features in libraries or in source code.

3. Percentage of pre written source code and source written by us.

Student ID	Student Name	Student signature
19103175	Khushi Jain	
19103070	Megha Agarwal	
19103071	Pavini Jain	



---

**Declaration by Supervisor (To be filled by Supervisor only)**

I **Dr. Parul Agarwal** declares that I above submitted project with Titled **Time Series Forecasting of Air Quality Prediction** was conducted in my supervision. The project is original and neither the project was copied from External sources not it was submitted earlier in JIIT. I authenticate this project.

(Any Remarks by Supervisor)

Signature (Supervisor)

---

## **1. Motivation behind making the project**

When it comes to human health, clean air is the most basic commodity. Today, poor air quality is one of the leading causes of a variety of severe health problems. To estimate the impact on our health, we must be aware of the air quality in our neighbourhood, city, and country. The intricate interaction of various components, including chemical reactions, climatic aspects, and emissions from natural and manmade sources, results in air quality.

The levels of air pollution in most urban areas have been a source of severe worry. People have the right to know the quality of the air they breathe. However, the data collected by the National Ambient Air Monitoring Network is reported in a format that is difficult to comprehend by the average person, and hence the current air quality information system does not support people's participation in air quality improvement activities.

The main goal of this project is to investigate the state and quality of the air by measuring the Air Quality Index (AQI) and comparing the measured values to standard values in order to create environmental impact.

Therefore, building a forecasting system for predicting the air quality based on the levels of concentration of individual pollutants and various meteorological parameters will be useful for the population's health.

---

## 2. Introduction

Contamination of the air, in particular in metropolitan areas, is a very well-known problem. The ever-growing population of cities and the increasing level of motorization contribute to the ever-increasing traffic volume, and consequently, the ever-increasing exhaust gases emissions. At the same time, the thickening of city buildings reduces ventilation and increases the porosity of surface, which ends up decreasing the effect of the wind on the evacuation of contamination. The typical sources of air pollution are well-known, but difficult to eliminate, at least completely. Thus, most studies are focused on determining the impact of factors that may modify the concentrations of contaminants in the atmosphere such as transformation, retention or evacuation.

In an attempt to make air quality measurement easier to understand, the ministry of environment and forests launched a National Air Quality Index (AQI). It will put out real time data about level of pollutants in the air and inform people about possible impacts on health.

The air quality index (AQI) is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health within a short time period. The purpose of the AQI is to help people know how the local air quality impacts their health. The Environmental Protection Agency (EPA) calculates the AQI for five major air pollutants, for which national air quality standards have been established to safeguard public health.

1. Ground-level ozone
2. Particle pollution/particulate matter (PM<sub>2.5</sub>/pm 10)
3. Carbon Monoxide
4. Sulphur dioxide
5. Nitrogen dioxide

The higher the AQI value, the greater the level of air pollution and the greater the health concerns. The concept of AQI has been widely used in many developed countries for over the last three decades. AQI quickly disseminates air quality information in real-time.

### **How is AQI calculated?**

India follows that the 500-point scale, wherein rating between 0 and 50 is considered good. Rating between 301 to 500 range is deemed hazardous. Every day monitors record concentrations of the major pollutants. These raw measurements are converted into a separate AQI value for each pollutant (ground-level ozone, particle pollution, carbon monoxide, and sulphur dioxide) using standard formulae developed by EPA. The highest of these AQI values are reported as the AQI value for that day.

---

## **Air Quality Index Categories:**

**Good (0–50)** - Minimal Impact

**Satisfactory (51–100)** - May cause minor breathing difficulties in sensitive people.

**Moderately polluted (101–200)** - May cause breathing difficulties in people with lung disease like asthma, and discomfort to people with heart disease, children and older adults.

**Poor (201–300)** - May cause breathing difficulties in people on prolonged exposure, and discomfort to people with heart disease

**Very Poor (301–400)** - May cause respiratory illness in people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases.

**Severe (401-500)** - May cause respiratory issues in healthy people, and serious health issues in people with lung/heart disease. Difficulties may be experienced even during light physical activity.

## **Objectives of Air Quality Index (AQI)**

- Comparing air quality conditions at different locations/cities.
- It also helps in identifying faulty standards and inadequate monitoring programmes.
- AQI helps in analysing the change in air quality (improvement or degradation).
- AQI informs the public about environmental conditions. It is especially useful for people suffering from illnesses aggravated or caused by air pollution.

## **3. Problem Statement**

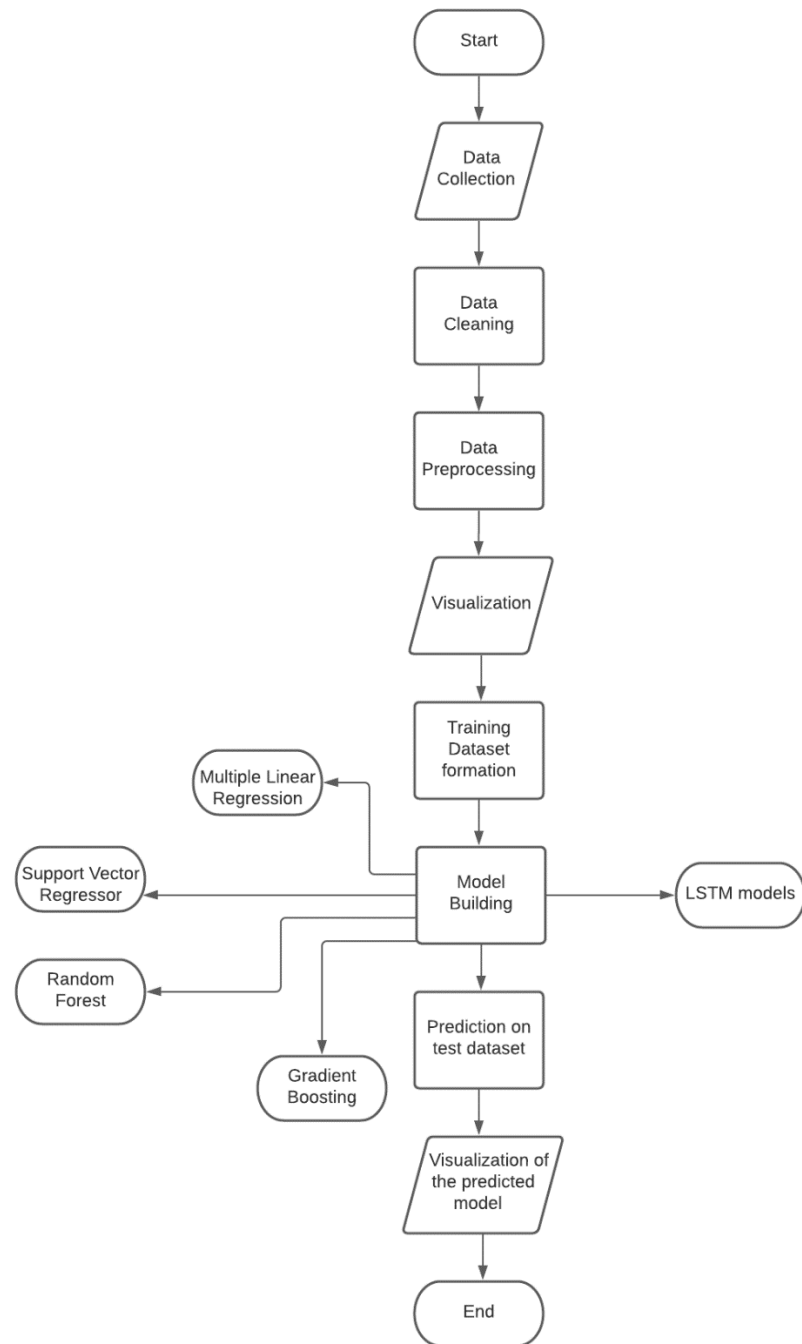
In this project our main aim is to develop an efficient approach for forecasting the air quality index of Bangalore using meteorological parameters and concentration of major air pollutants using various Machine Learning Algorithms and Deep Neural Networks like LSTM.

## **4. Methodology**

1. Data Collection
2. Data Cleaning
3. Data Pre-processing
4. Visualization
5. Training Dataset Formation
6. Model Building
7. Prediction on testing dataset
8. Visualization of predicted model

---

## 4.1. Flow chart



**Figure-4.1**

---

## 4.2. Data Collection

Meteorological parameters like Maximum temperature (°C), Minimum temperature (°C), Minimum temperature (°C), Average relative humidity (%), Total rainfall and / or snowmelt (mm), Average visibility (Km), Average wind speed (Km/h), Maximum sustained wind speed (Km/h) were taken from [TuTiempo.net](https://www.tutiempo.net).

Concentration of various Air Pollutants like PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3 were used. Data was collected from [Kaggle](https://www.kaggle.com)

## 4.3. Data Cleansing

Data cleansing entails identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Meteorological parameters were web-scraped from WEBSITE and the air pollutants concentrations were taken from Kaggle dataset.

From the data, various features which are not useful in AQI prediction like 'NH3', 'NO', 'Benzene', 'Toluene', 'Xylene', 'AQI Bucket' are removed. Finally, the independent features used are PM2.5, PM10, NO2, NOx, CO, SO2, O3, average temp, maximum temperature, minimum temperature, humidity, visibility, and wind speed.

Rows with missing AQI values were dropped whereas the missing values in independent features were interpolated.

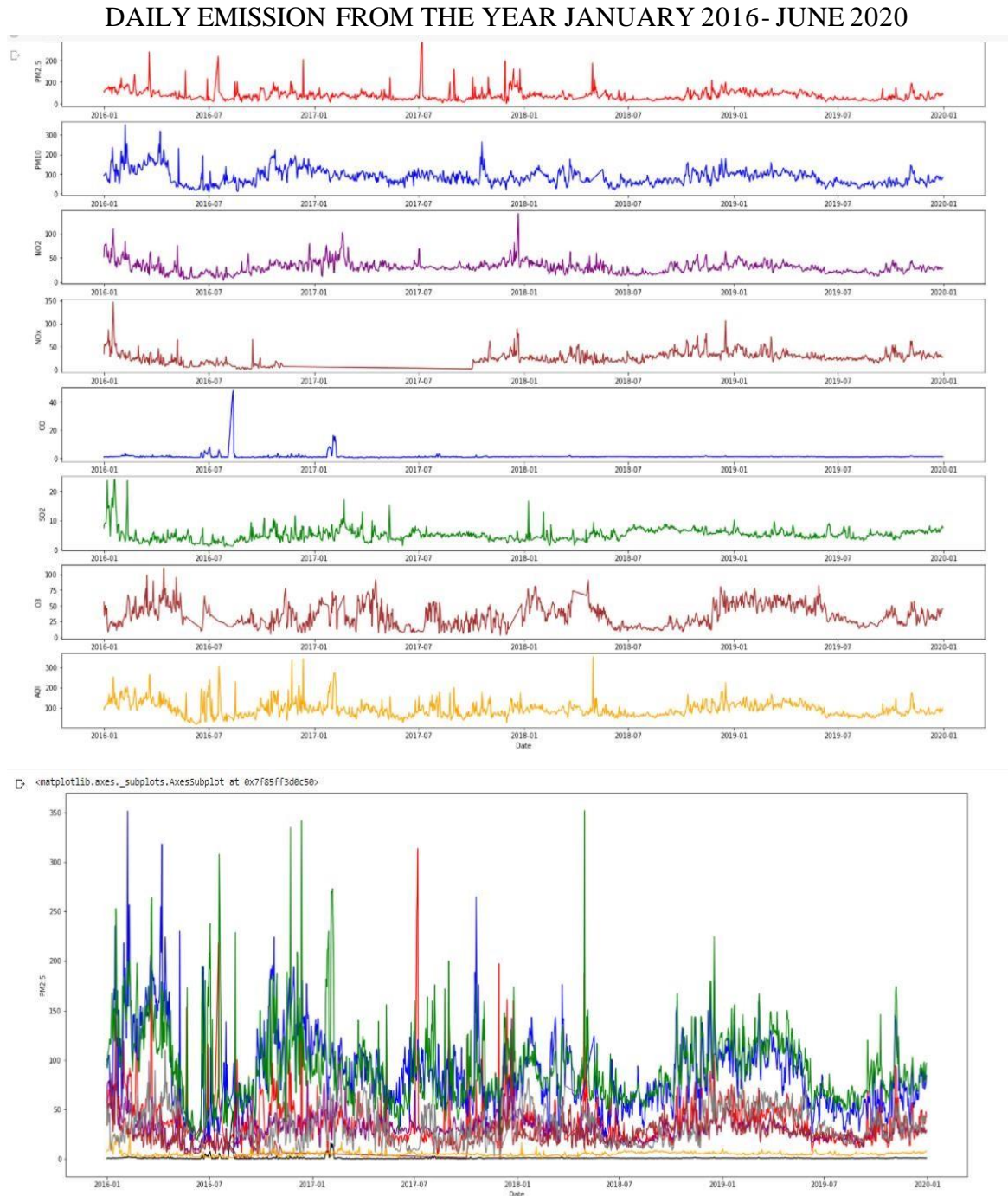
Feature Scaling was performed using Standard Scalar.



Figure-4.2

## 4.4. Data Visualization

The data obtained after cleansing was then visualized to check whether the data is proper or not.



**Figure-4.3**



## MONTHLY AVERAGE EMISSION FROM THE YEAR JANUARY 2016- JUNE 2020

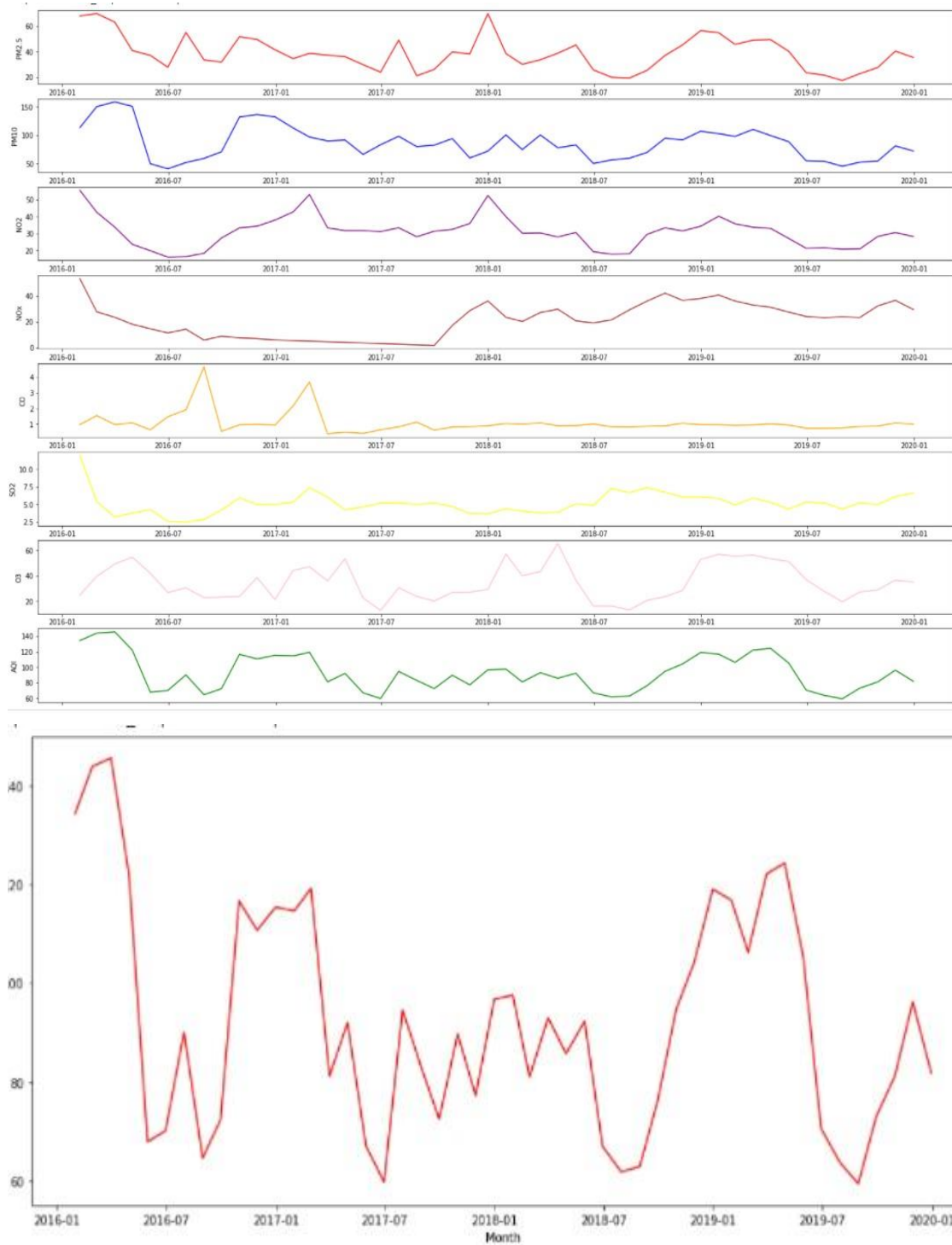


Figure-4.4

## MAXIMUM AVERAGE EMISSION FROM THE YEAR JANUARY 2016- JUNE 2020

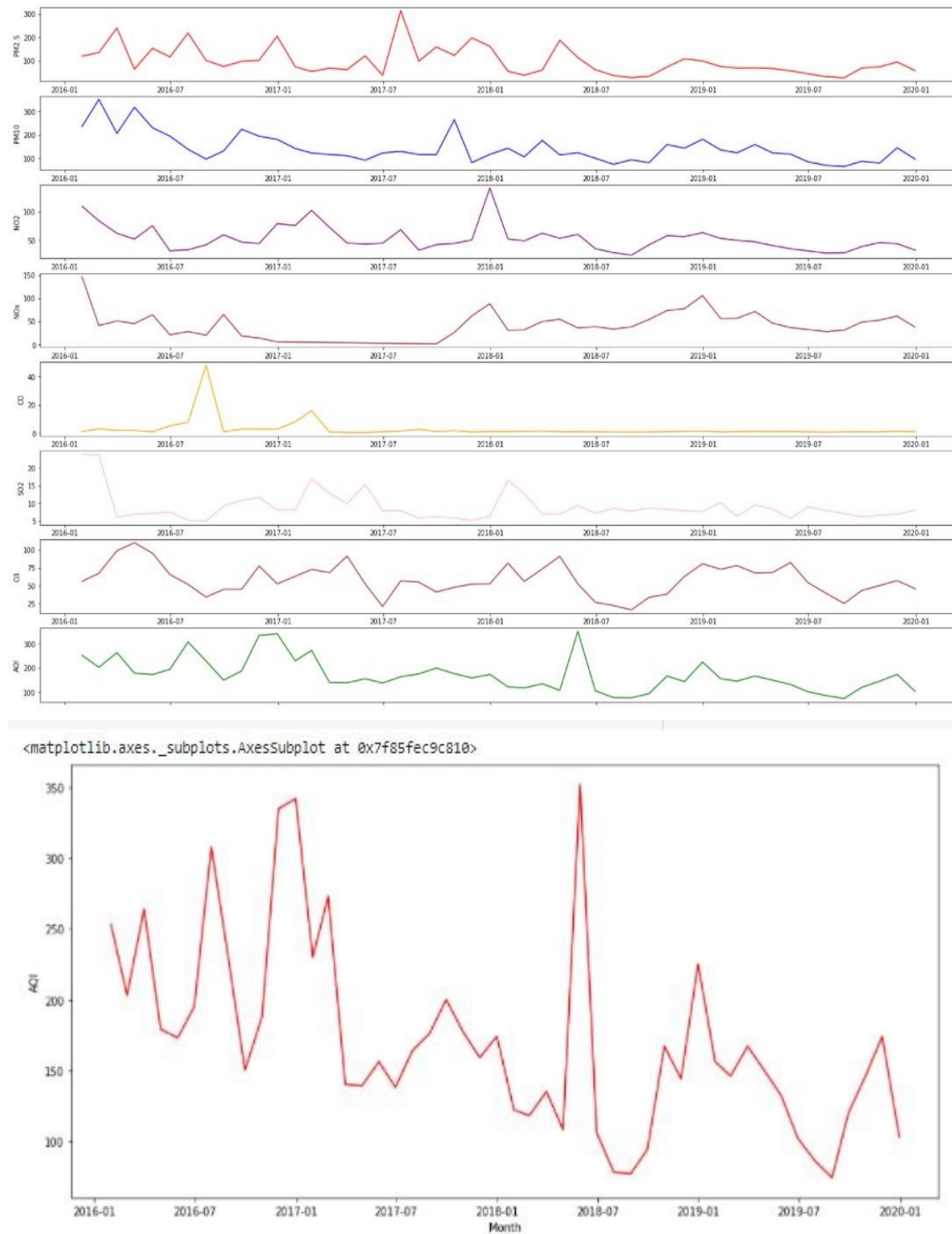


Figure-4.5

---

## **4.5. Data Preprocessing**

The first and most important criterion to ensure the effective construction of forecasting models is data quality and representativity. The capacity of a machine learning algorithm to generalise is often influenced by the data preparation phase. Missing data imputation, eliminating or changing outlier observations, data transformation (typically normalisation and standardisation), and feature engineering are all examples of data preparation. While the first two phases are important for obtaining more precise and full data sets, the third step is often used to obtain data that is more consistently distributed and to reduce data variability. Finally, the fourth phase is utilised to generate a new dataset that is often smaller and more informative. Feature extraction and feature selection are usually included in the final stage.

## **4.6. Training Dataset Formation**

Data is collected for the years 2016-2020, out of which data from 2016-2019 will be used for training the model and the collected AQI values for 2020 will be compared with the values forecasted using the LSTM model and various other machine learning models as well.

## **4.7. Architecture**

### **4.7.1. MLR**

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

### **4.7.2. SVR**

The supervised learning algorithm Support Vector Regression is used to predict discrete values. SVMs and Support Vector Regression are both based on the same premise. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

---

### 4.7.3. RANDOM FOREST

A random forest is a machine learning technique for solving classification and regression problems. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers. A random forest algorithm consists of many decision trees. The ‘forest’ generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

### 4.7.4. GRADIENT BOOSTING

Gradient boosting is a sort of boosting used in machine learning. It is based on the assumption that when the best potential next model is coupled with prior models, the overall prediction error is minimised. To decrease error, the fundamental notion is to specify the target outcomes for the next model.

### 4.7.5. RNN

Recurrent Neural Network (RNN) is a type of Neural Network where the output from previous steps are fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence. RNN has a “memory” which remembers all information about what has been calculated. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity of parameters, unlike other neural networks.

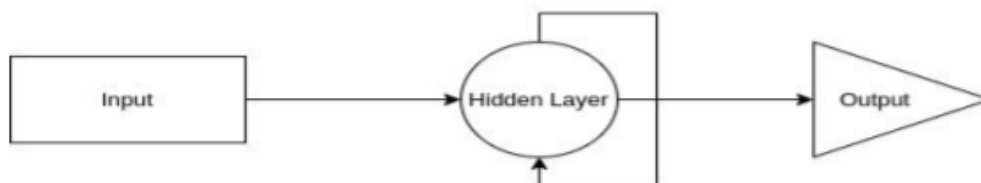


Figure 4.6

---

#### 4.7.6 LSTM

Long Short-Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochester & Schmid Huber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting and classifying on the basis of time series data.

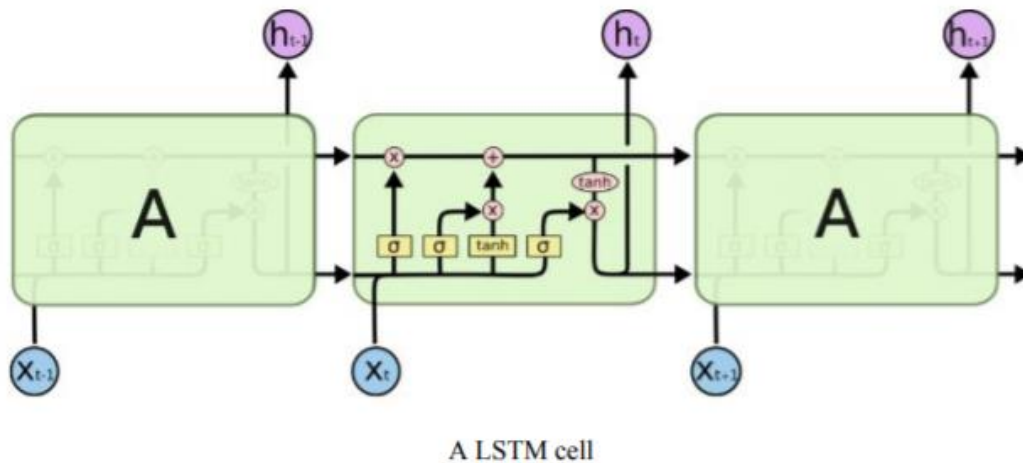


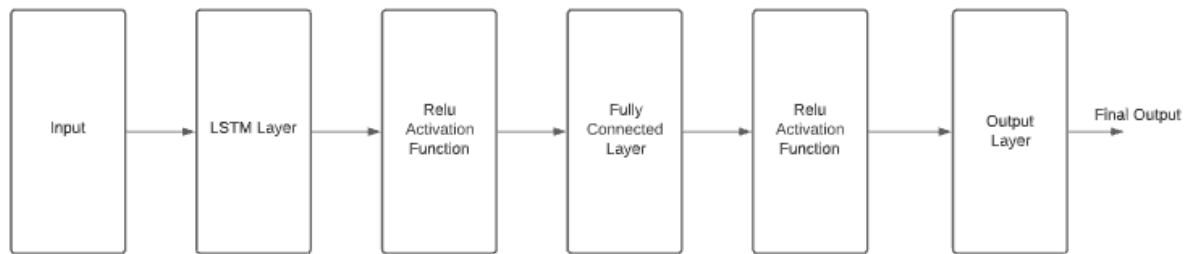
Figure 4.7

#### 4.7.7. Final Proposed Architecture

The given dataset is converted in a way that a row contains the data for previous seven days and the target feature is the AQI value for the eight day which means that the AQI values for the last seven days will be used to forecast the AQI for next day.

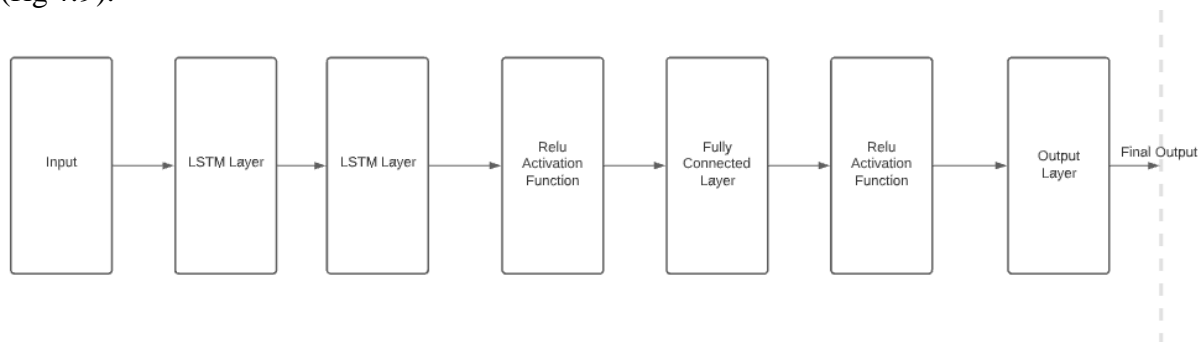
We developed four models containing LSTM layers and fully connected layers. The output from the LSTM and fully connected layer is passed through the ReLU activation function. Adam optimizer is used to optimize the weights and learning rate values. We used Mean Squared Error as the loss function. An initial value of 0.001 is used as the learning rate.

First model comprises of one LSTM layer, one fully connected layer and an output layer. Output from the LSTM and fully connected layer is passed through the ReLU activation function (fig 4.8).



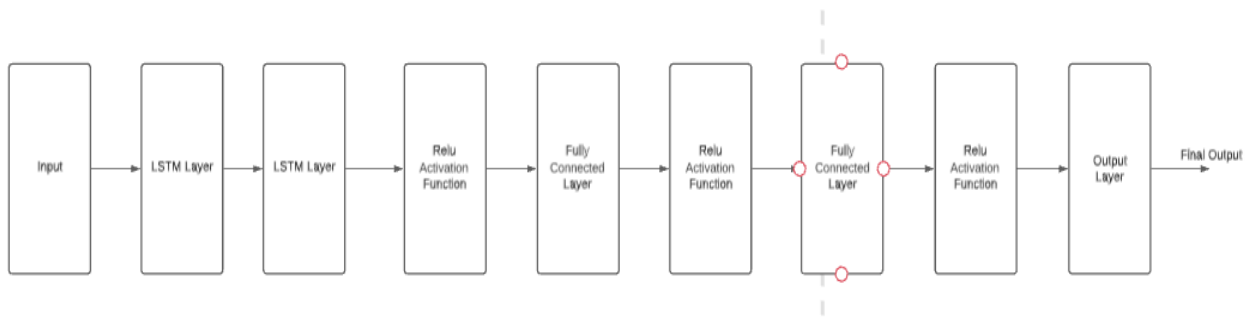
**Figure 4.8**

Second model comprises of two LSTM layers, one fully connected layer and an output layer. The output from the LSTM and fully connected layer is passed through the ReLU activation function (fig 4.9).



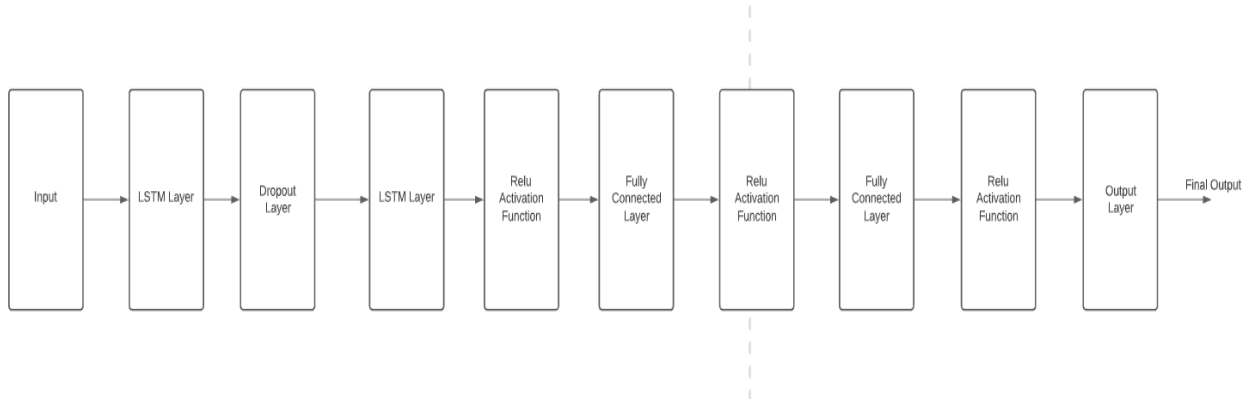
**Figure 4.9**

Third model comprises of two LSTM layers, two fully connected layers and an output layer. Output from the layers is passed through ReLU (fig 4.10).



**Figure 4.10**

Finally, fourth and the final model is developed by adding a dropout layer with a dropout rate 0.5 to the third model. This model gave the best forecasting results for AQI value. The arrangement of the models is shown in fig. 4.11.



**Figure 4.11**

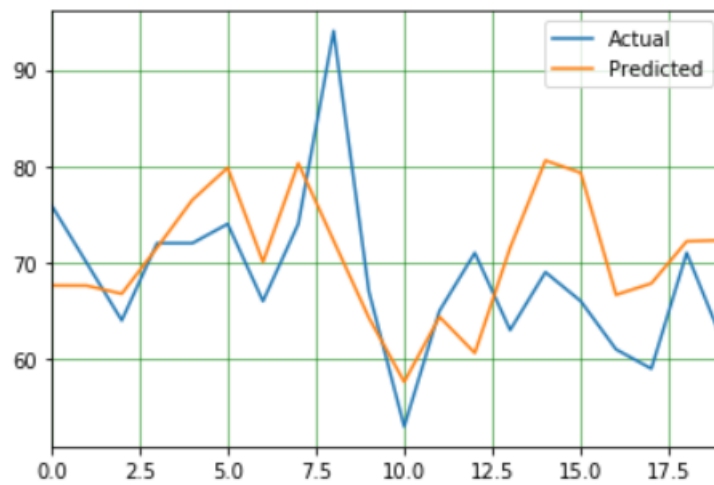
## 5. Results

Now, we have compared the actual and the predicted AQI values obtained from the above models using a line plot.

### 5.1. Predictions

#### 5.1.1. Prediction by Multiple Linear Regression

The AQI values for the testing dataset was predicted with a R2 Score of 0.817020 and RMS Error of 0.19739.



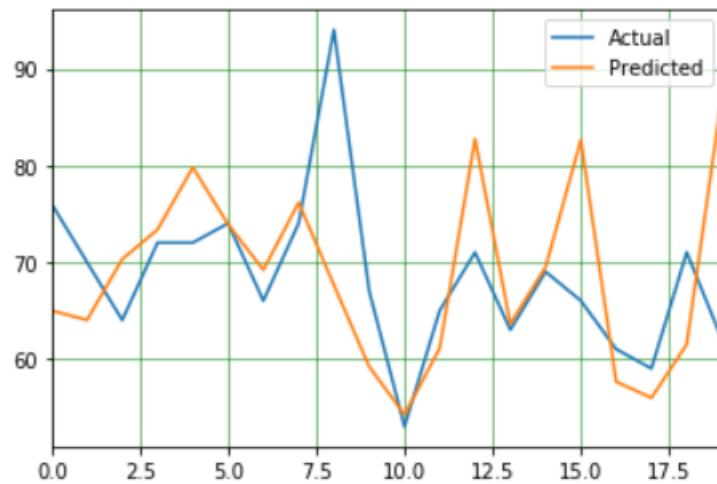
**Figure-5.1**

	ACTUAL AQI	PREDICTED AQI	DIFFERENCE	PERCENTAGE ERROR
0	82.0	86.648713	-4.648713	5.669162
1	81.0	86.326485	-5.326485	6.575908
2	85.0	82.594906	2.405094	2.829523
3	95.0	116.165523	-21.165523	22.279498
4	118.0	111.206382	6.793618	5.757303
5	81.0	76.090810	4.909190	6.060728
6	75.0	78.156551	-3.156551	4.208734
7	93.0	93.544459	-0.544459	0.585440
8	101.0	101.126981	-0.126981	0.125724
9	94.0	98.016975	-4.016975	4.273377
10	78.0	84.779536	-6.779536	8.691713
11	79.0	93.472071	-14.472071	18.319077
12	106.0	109.544582	-3.544582	3.343946
13	121.0	120.977642	0.022358	0.018477
14	111.0	105.294744	5.705256	5.139870

**Table-5.1**

### 5.1.2. Prediction by Support Vector Regression

The AQI values for the testing dataset was predicted with a R2 Score of 0.80630 and RMS Error of 0.22212.



**Figure-5.2**

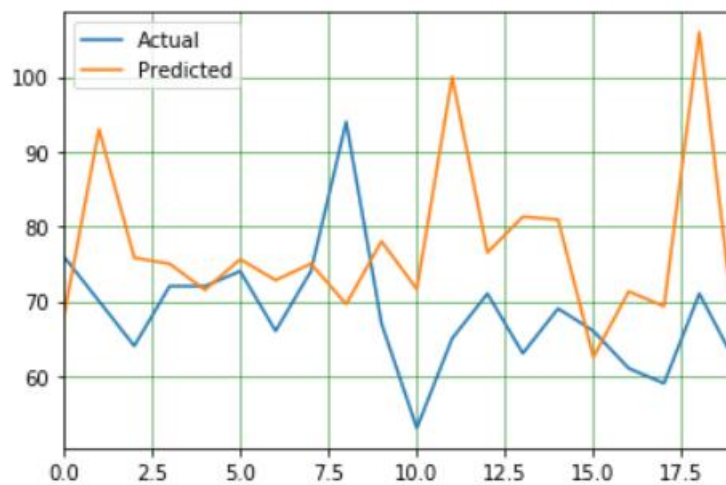


	ACTUAL AQI	PREDICTED AQI	DIFFERENCE	PERCENTAGE ERROR
0	82.0	76.274254	5.725746	6.982617
1	81.0	80.203947	0.796053	0.982782
2	85.0	78.245953	6.754047	7.945938
3	95.0	114.283815	-19.283815	20.298753
4	118.0	106.241231	11.758769	9.965058
5	81.0	76.158900	4.841100	5.976667
6	75.0	70.661163	4.338837	5.785116
7	93.0	82.520098	10.479902	11.268712
8	101.0	93.152173	7.847827	7.770126
9	94.0	88.308799	5.691201	6.054469
10	78.0	80.321475	-2.321475	2.976250
11	79.0	84.183218	-5.183218	6.561036
12	106.0	108.131829	-2.131829	2.011160
13	121.0	119.507700	1.492300	1.233306
14	111.0	101.134023	9.865977	8.888267

**Table-5.2**

### 5.1.3. Prediction by Random Forest

The AQI values for the testing dataset was predicted with a R2 Score of 0.757429 and RMS Error of 0.248566.



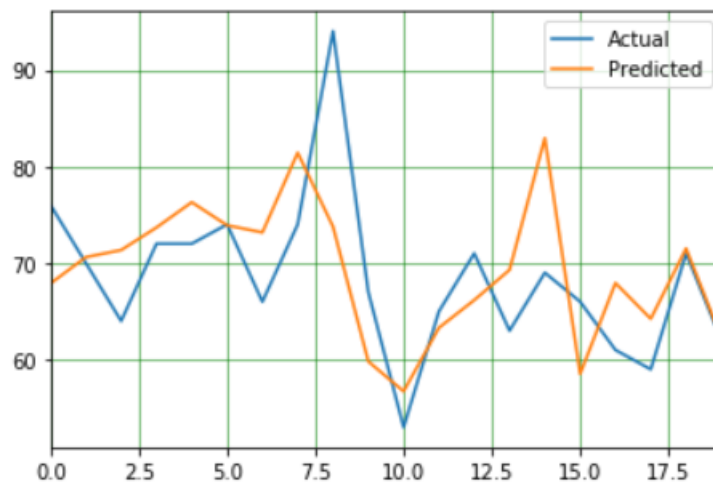
**Figure-5.3**

	ACTUAL AQI	PREDICTED AQI	DIFFERENCE	PERCENTAGE ERROR
0	82.0	83.2	-1.2	1.463415
1	81.0	97.7	-16.7	20.617284
2	85.0	79.1	5.9	6.941176
3	95.0	108.2	-13.2	13.894737
4	118.0	104.6	13.4	11.355932
5	81.0	74.5	6.5	8.024691
6	75.0	78.8	-3.8	5.066667
7	93.0	88.0	5.0	5.376344
8	101.0	92.6	8.4	8.316832
9	94.0	92.3	1.7	1.808511
10	78.0	79.7	-1.7	2.179487
11	79.0	82.8	-3.8	4.810127
12	106.0	110.4	-4.4	4.150943
13	121.0	122.8	-1.8	1.487603
14	111.0	88.8	22.2	20.000000

**Table-5.3**

#### 5.1.4. Prediction by Gradient Boosting

The AQI values for the testing dataset was predicted with a R2 Score of 0.78394 and RMS Error of 0.234586.

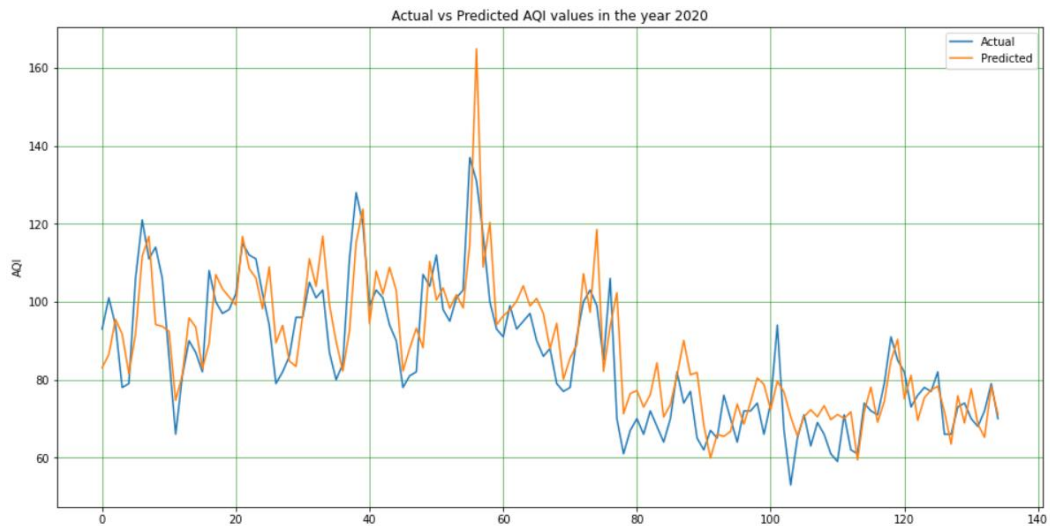


**Figure-5.4**

	ACTUAL AQI	PREDICTED AQI	DIFFERNCE	PERCENTAGE ERROR
0	82.0	80.832990	1.167010	1.423183
1	81.0	93.598285	-12.598285	15.553439
2	85.0	75.156039	9.843961	11.581131
3	95.0	114.260451	-19.260451	20.274159
4	118.0	107.509410	10.490590	8.890331
5	81.0	79.698389	1.301611	1.606927
6	75.0	74.979144	0.020856	0.027808
7	93.0	87.291481	5.708519	6.138193
8	101.0	100.737800	0.262200	0.259604
9	94.0	94.906258	-0.906258	0.964104
10	78.0	75.753040	2.246960	2.880717
11	79.0	90.596251	-11.596251	14.678798
12	106.0	108.012470	-2.012470	1.898556
13	121.0	127.385603	-6.385603	5.277358
14	111.0	102.076964	8.923036	8.038772

**Table-5.4**

### 5.1.5. Prediction by LSTM MODEL 1

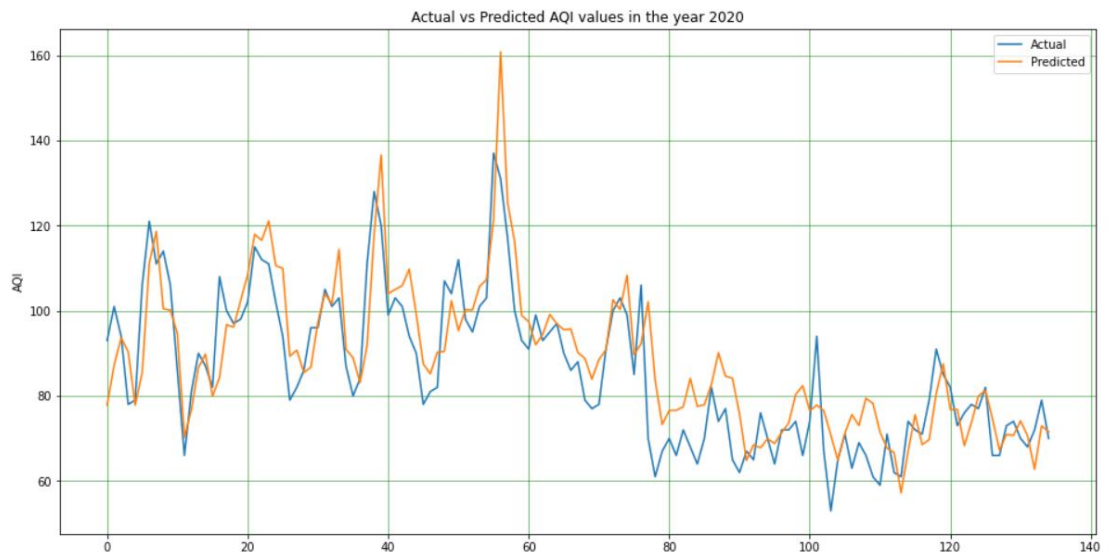


**Figure-5.5**

	ACTUAL AQI	PREDICTED AQI	DIFFERENCE	PERCENTAGE ERROR
0	93.0	83.086464	9.913536	10.659717
1	101.0	86.443192	14.556808	14.412682
2	94.0	95.473206	-1.473206	1.567240
3	78.0	91.716141	-13.716141	17.584797
4	79.0	81.480530	-2.480530	3.139911
5	106.0	91.770256	14.229744	13.424286
6	121.0	111.694244	9.305756	7.690707
7	111.0	116.772820	-5.772820	5.200738
8	114.0	94.161079	19.838921	17.402563
9	106.0	93.667404	12.332596	11.634524
10	86.0	92.440254	-6.440254	7.488668
11	66.0	74.703308	-8.703308	13.186831
12	81.0	81.210060	-0.210060	0.259333
13	90.0	95.833717	-5.833717	6.481908
14	87.0	93.519287	-6.519287	7.493433

**Table-5.5**

### 5.1.6. Prediction by LSTM MODEL 2

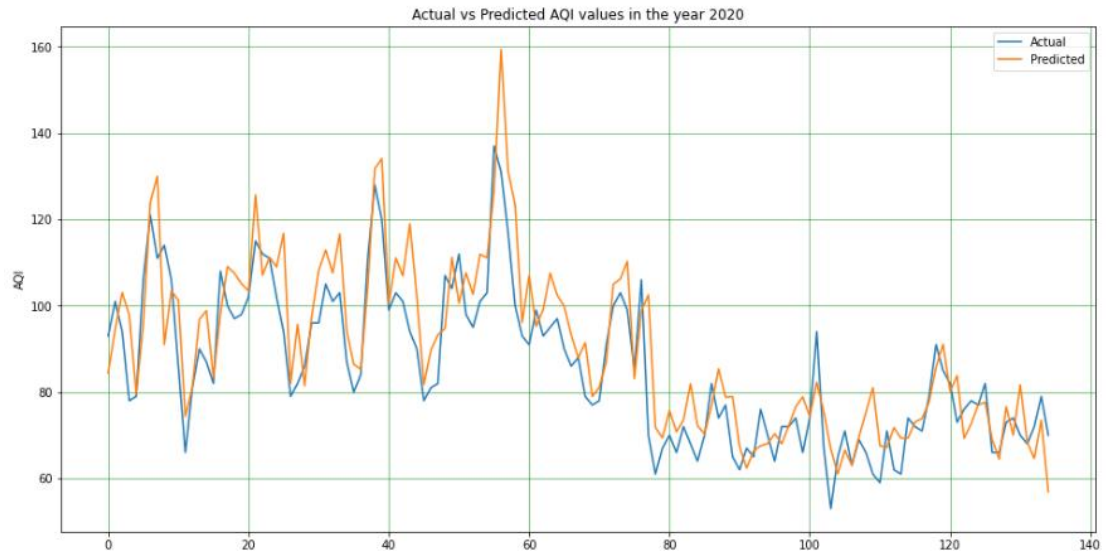


**Figure-5.6**

	ACTUAL AQI	PREDICTED AQI	DIFFERENCE	PERCENTAGE ERROR
0	93.0	77.840927	15.159073	16.300077
1	101.0	87.173660	13.826340	13.689445
2	94.0	93.632301	0.367699	0.391169
3	78.0	90.343864	-12.343864	15.825467
4	79.0	77.820915	1.179085	1.492512
5	106.0	85.494705	20.505295	19.344618
6	121.0	111.067261	9.932739	8.208876
7	111.0	118.636559	-7.636559	6.879783
8	114.0	100.442902	13.557098	11.892191
9	106.0	100.138901	5.861099	5.529339
10	86.0	94.637344	-8.637344	10.043424
11	66.0	70.281250	-4.281250	6.486742
12	81.0	76.716568	4.283432	5.288188
13	90.0	86.680717	3.319283	3.688092
14	87.0	89.800369	-2.800369	3.218815

**Table-5.6**

### 5.1.7. Prediction by LSTM MODEL 3

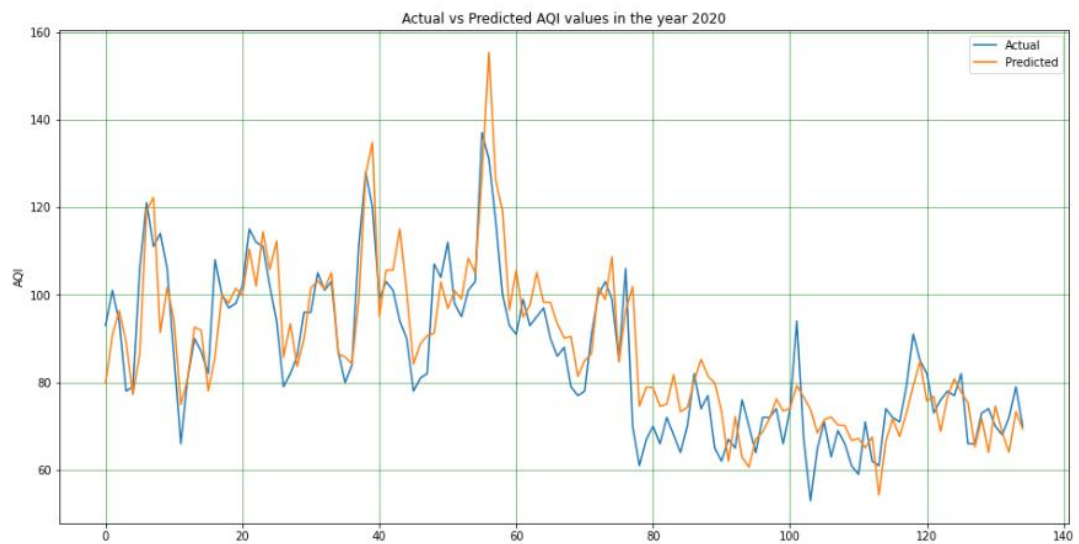


**Figure-5.7**

	ACTUAL AQI	PREDICTED AQI	DIFFERENCE	PERCENTAGE ERROR
0	93.0	84.404427	8.595573	9.242552
1	101.0	94.565178	6.434822	6.371111
2	94.0	103.037682	-9.037682	9.614554
3	78.0	97.845886	-19.845886	25.443443
4	79.0	79.680466	-0.680466	0.861349
5	106.0	95.541817	10.458183	9.866211
6	121.0	123.937798	-2.937798	2.427932
7	111.0	129.995743	-18.995743	17.113281
8	114.0	90.975487	23.024513	20.196941
9	106.0	103.436035	2.563965	2.418835
10	86.0	101.273567	-15.273567	17.759962
11	66.0	74.369637	-8.369637	12.681268
12	81.0	81.142883	-0.142883	0.176399
13	90.0	96.829712	-6.829712	7.588569
14	87.0	98.851814	-11.851814	13.622776

**Table-5.7**

#### 5.1.8. Prediction by LSTM MODEL 4



**Figure-5.8**

---

	ACTUAL AQI	PREDICTED AQI	DIFFERENCE	PERCENTAGE ERROR
0	93.0	79.834099	13.165901	14.156882
1	101.0	90.716225	10.283775	10.181956
2	94.0	96.306183	-2.306183	2.453386
3	78.0	89.291634	-11.291634	14.476453
4	79.0	77.264694	1.735306	2.196590
5	106.0	86.610535	19.389465	18.291948
6	121.0	119.238716	1.761284	1.455606
7	111.0	122.201530	-11.201530	10.091469
8	114.0	91.420525	22.579475	19.806559
9	106.0	101.708183	4.291817	4.048883
10	86.0	94.254997	-8.254997	9.598834
11	66.0	74.919205	-8.919205	13.513947
12	81.0	80.816925	0.183075	0.226018
13	90.0	92.619576	-2.619576	2.910640
14	87.0	91.858055	-4.858055	5.583972

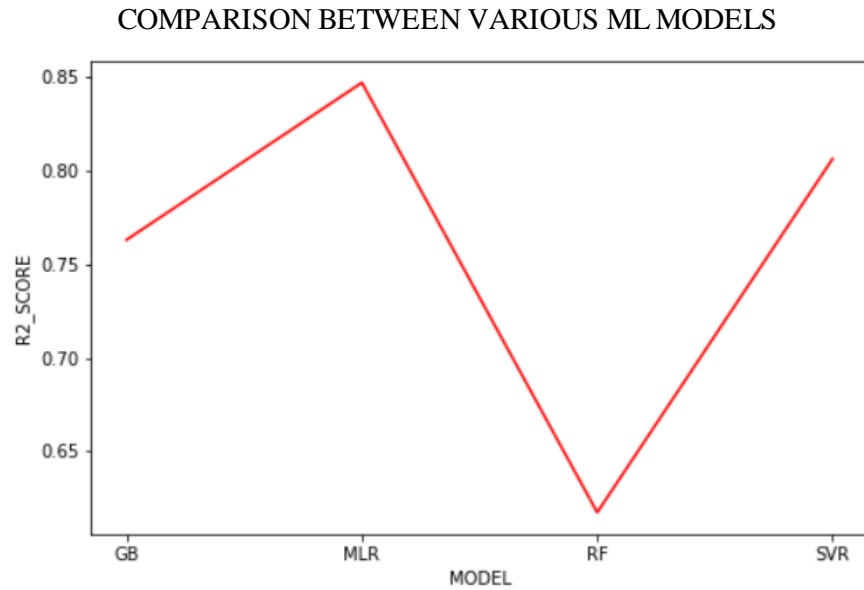
**Table-5.8**

## 5.2. Errors and Losses

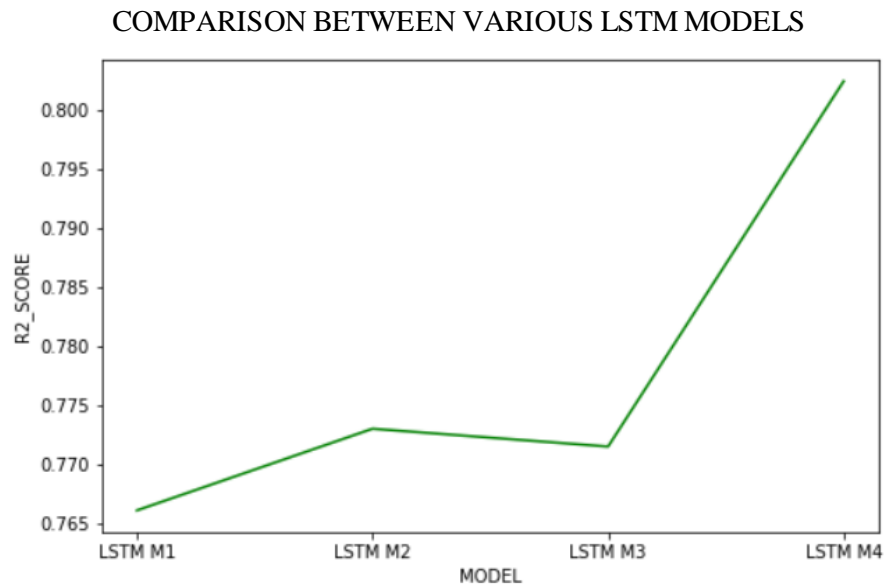
Comparison of different models based on their R2 score, Mean absolute error and root mean squared error is shown below:

### 5.2.1. R2 Squared

It is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model. R-squared explains to what extent the variance of one variable explains the variance of the second variable.



**Figure-5.9**

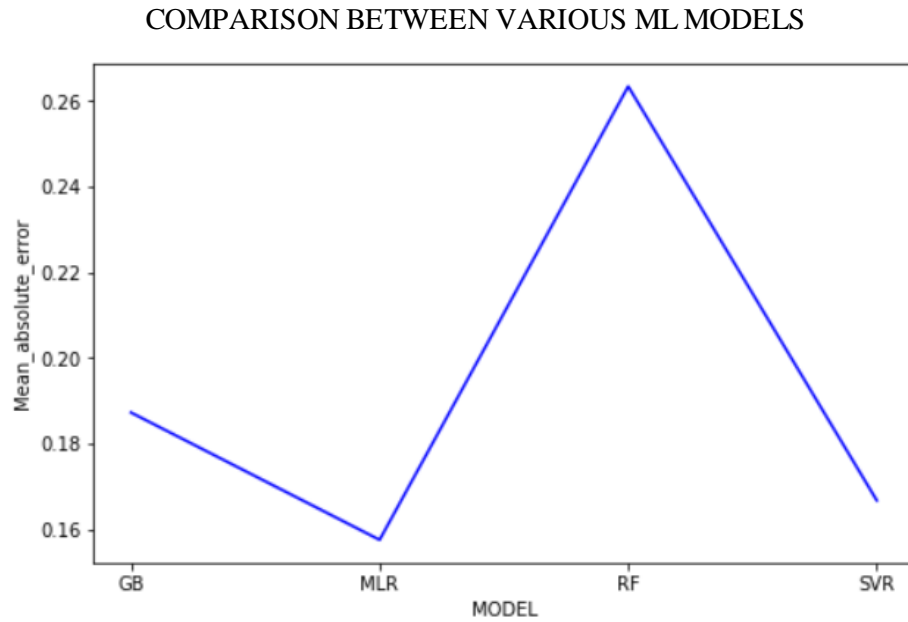


**Figure-5.10**

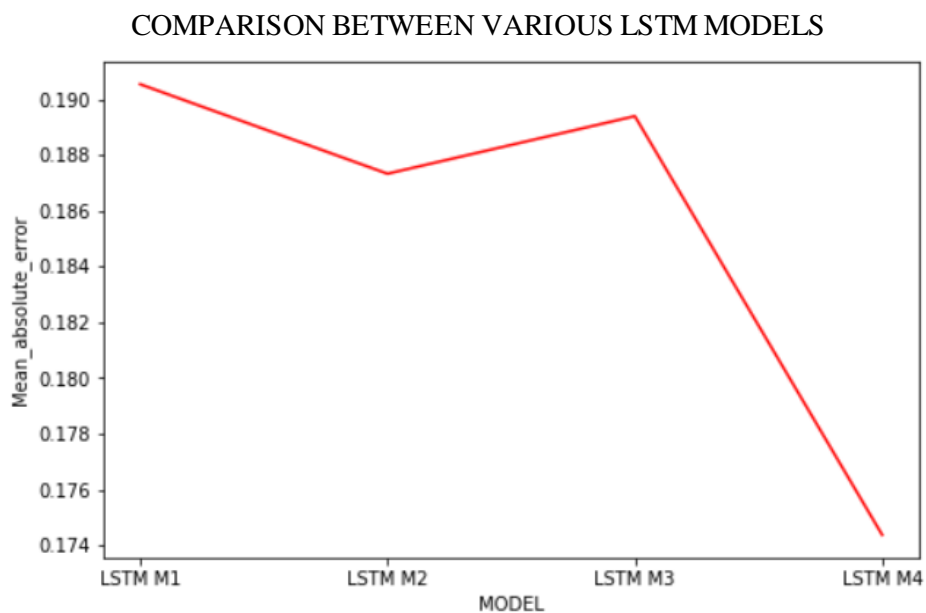
### 5.2.2. Mean Absolute Error

The magnitude of the difference between the individual measurement and the true value of the quantity is called the absolute error of the measurement. The arithmetic mean of all the absolute error is taken as the mean absolute error of the value of the physical quantity.





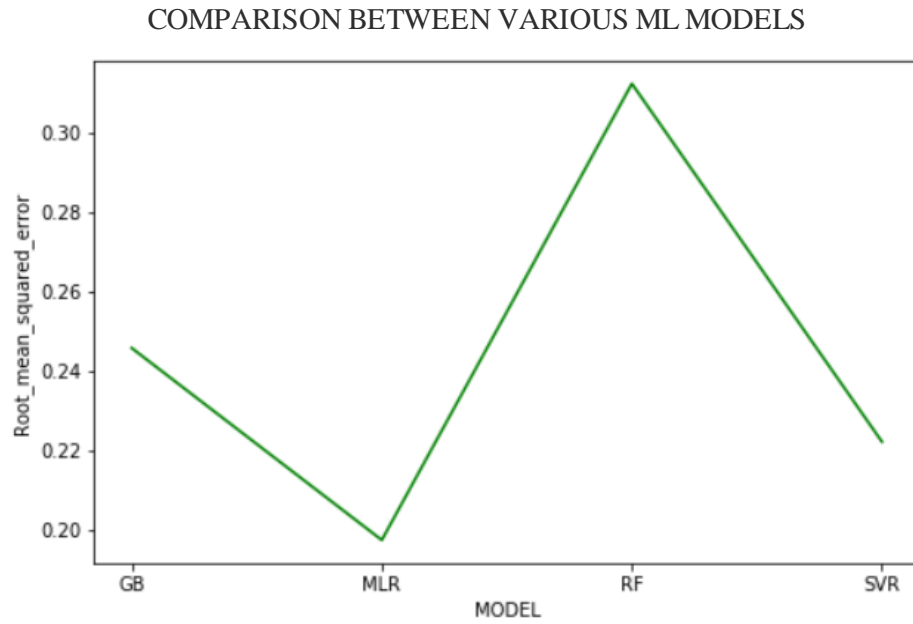
**Figure-5.11**



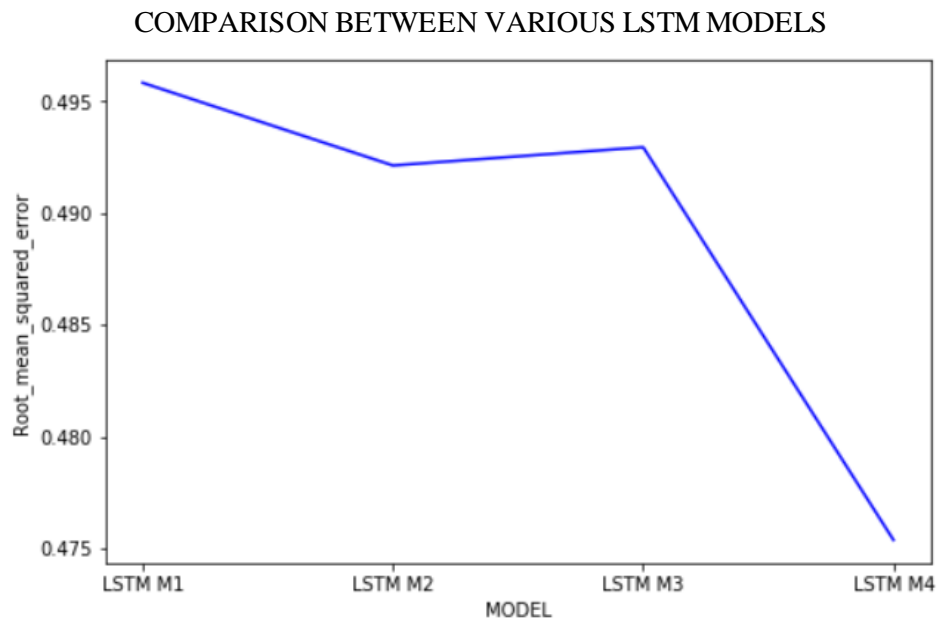
**Figure-5.12**

### 5.2.3. Root Mean Squared Error

It is the square root of the mean of the square of all of the error. RMSE is a good measure of accuracy, but only to compare prediction errors of different models or model configurations for a particular variable and not between variables, as it is scale-dependent.



**Figure-5.13**



**Figure-5.14**

---

### 5.3. Comparison

#### 5.3.1. Comparison between various ML Models

MODEL	R2_SCORE	Mean absolute error	Root mean squared error
MULTIPLE LINEAR REGRESSION	0.81702036287260	0.15755965269436	0.19739750373716397
SUPPORT VECTOR REGRESSION	0.80630010071042	0.16675010409400	0.22212095937096654
RANDOM FOREST	0.75742992832956	0.19775706724581	0.24856698890550402
GRADIENT BOOSTING	0.78394965329306	0.17724614948669	0.23458612795481937

**Table-5.9**

#### 5.3.2. Comparison between various LSTM Models

MODEL	R2_SCORE	Mean absolute error	Root mean squared error
LSTM MODEL 1	0.766110725303	0.19053435	0.49583682
LSTM MODEL 2	0.773016977716	0.18732437	0.49213535
LSTM MODEL 3	0.771508592511	0.18938948	0.49295092
LSTM MODEL 4	0.802407929275	0.1743705	0.4753665

**Table-5.10**

---

## 6. Conclusion

The air quality index (AQI) or air pollution index (API) is a standard method of informing the public about the severity of air pollution. Various researchers/environmental agencies have created a number of ways for determining AQI or API in the past, but there is no globally approved method that is adequate for all scenarios. In computing the AQI or API, different methods utilise different aggregation functions and take into account different types and amounts of contaminants.

When dangerous or excessive quantities of specific chemicals such as gases, particles, and biological molecules are introduced into the atmosphere, it is referred to as air pollution. Excessive emissions have apparent repercussions, such as disease and mortality among populations and other living species, as well as crop damage.

Because of the dynamic nature, volatility, and great unpredictability in location and time of pollutants and particles, predicting air quality is a difficult undertaking. Simultaneously, due to the recognised significant repercussions of air pollution on humans and the environment, the ability to model, predict, and monitor air quality is becoming increasingly vital, particularly in metropolitan areas.

---

## 7. References

1. <https://www.sciencedirect.com/science/article/pii/S1309104215304700>
2. <https://downloads.hindawi.com/journals/jcse/2012/518032.pdf>
3. <https://www.iosrjournals.org/iosr-jestft/papers/vol3-issue5/A0350108.pdf>
4. <https://www.mdpi.com/1424-8220/16/1/86>
5. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179763#sec002>
6. <https://www.hindawi.com/journals/amete/2018/3506394/>
7. <http://www.mecs-press.net/ijisa/ijisa-v111-n2/IJISA-V111-N2-3.pdf>
8. <https://www.hindawi.com/journals/complexity/2020/8049504/>
9. <https://www.hindawi.com/journals/wcmc/2021/9627776/>
10. <https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html>
11. <https://pytorch.org/>
12. <https://www.ijert.org/research/a-novel-air-quality-prediction-model-using-artificial-neural-networks-IJERTV3IS21323.pdf>