AI60002

Machine Learning For Earth System Science

Term Project

Air Quality Forecasting using Neural Networks

Sachin Khoja - 18NA30015

Table of Content

Introduction

Air pollution is the introduction of particulates, biological molecules, or other harmful substances into the Earth's atmosphere. It is a global concern and a major environmental health problem. It is the fifth leading risk factor for mortality worldwide. People are dying more because of air pollution than malnutrition, road traffic injuries, and alcohol use . According to the World Health Organization (WHO), 9 out of 10 people around the world breathe polluted air. Every year, around 7 million people die from exposure to air pollution . One-third of deaths from heart disease, lung cancer, and stroke are due to air pollution.

Due to rapid urbanization, many environmental hazards took place in the 20th century, including rise in air pollution levels. Air pollution is the introduction of chemicals, particulate matter, or biological matter that cause harm or discomfort to living organisms or damage the natural environment or atmosphere. Air pollutants are tiny and light particles and thus they stay in the atmosphere for a long duration and also easily bypass the filters of the human nose and throat due to their small size. According to a recent survey, the presence of particulate matter has caused 4.2 million deaths. Major air pollutants like Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Particulate Matter (PM2.5, PM10) and Ozone(O3) have drastic effects on human health. Thus, predicting the air quality has become a major concern. Particle size is critical in determining the particle deposition location in the human respiratory system. PM2.5, referring to particles with a diameter less than or equal to 2.5 μm, has been an increasing concern, as these particles can be deposited into the alveoli- the lung gas exchange region. The U.S. EPA revised the annual standard of PM2.5 by lowering the concentration to 12 μg/m3 to provide improved protection against health effects associated with long- and short-term exposure. Increased mortality and morbidity rates have been found in association with increased air pollutants. Thus, we considered PM2.5 as the label for classification. Meteorological data is critical in determining air pollutant consideration. The meteorological parameters considered by our model include: temperature, wind, relative humidity, dew point and pressure. Temperature affects air quality because of temperature inversion: the warm air above cooler air acts like a lid, suppressing vertical mixing and trapping the cooler air at the surface. As pollutants from vehicles, fireplaces, and industry are emitted into the air, the inversion traps these pollutants near the ground. Wind speed plays a big role in diluting pollutants. Generally, strong winds disperse pollutants, whereas light winds generally result in stagnant conditions allowing pollutants to build up over an area. Humidity could affect the diffusion of contaminants. Dew point indicates the amount of the moisture in the air. The higher the dew point the higher the moisture content in the air at a given temperature. Dew point and the concentration of the air pollutants are inversely proportional. Pressure is also inversely proportional to the air quality.

Aim

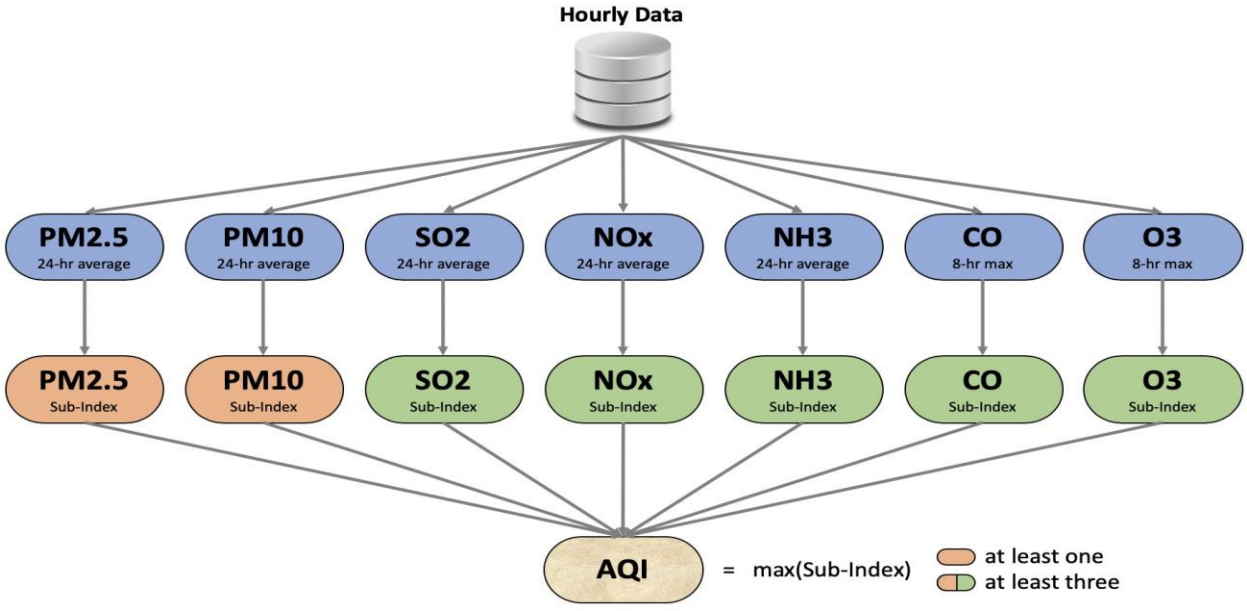
The Air Quality Index prediction problem has been in limelight in recent years, due to rapid increase in air pollution. A lot of work has been done in this field to take appropriate measures to maintain the air quality.In all these years, conventional approaches are used for ambient air quality assessments. Manual analysis of raw data is carried out in these approaches.However, these methods are inefficient, complex, and provide limited accuracy. With recent advancement in technology and research, novel air quality assessment techniques have been modeled We will discuss a deep-learning based approach to predict ambient air quality in Jodhpur, India.

The main contributions of this report are:

- LSTM model, proposed to predict air pollutants' concentration in Jodhpur; ●
  AQI calculation, to forecast the ambient air quality for the next hour;

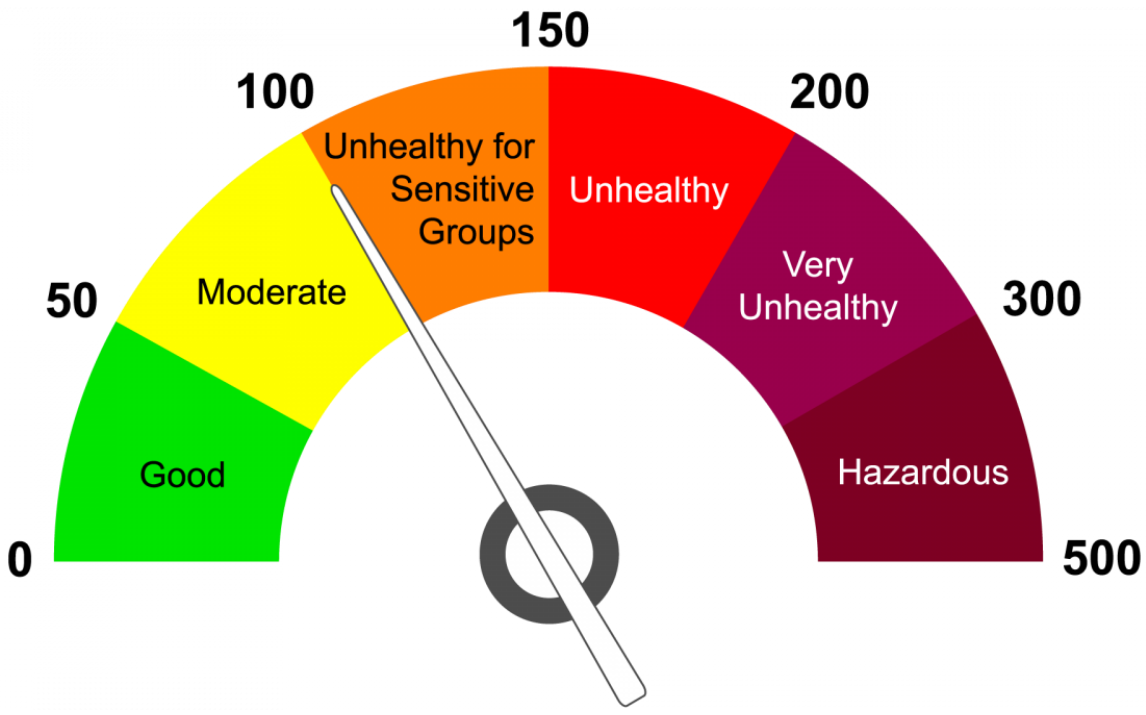
AQI

The quality of life in a place is measured using several factors in which air quality plays a vital role. Its measurement is based on the concentration of pollutants in the atmosphere and is called AQI. AQI is a method that transforms the weighted values of individual air pollution-related parameters (for example, pollutant concentrations) into a single number or set of numbers. In the AQI system, specific concentration ranges are grouped into air quality descriptor categories.In Indian AQI System (IND-AQI), the following eight pollutants are considered for calculation of AQI: CO, NO2, SO2, PM2.5, PM10, O3, NH3, and Pb. To present the status of air quality and its effects on human health, the following six air quality description categories have been adopted: Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe



- The AQI calculation uses 7 measures: **PM2.5, PM10, SO2, NOx, NH3, CO and O3**.
- For **PM2.5, PM10, SO2, NOx and NH3** the average value in the last 24-hrs is used with the condition of having at least 16 values.
- For **CO and O3** the maximum value in the last 8-hrs is used.
- Each measure is converted into a Sub-Index based on pre-defined groups.
- Sometimes measures are not available due to lack of measuring or lack of required data points.
- Final AQI is the maximum Sub-Index with the condition that at least one of PM2.5 and PM10 should be available and at least three out of the seven should be available.

The final AQI is the maximum Sub-Index among the available sub-indices with the condition that at least one of PM2.5 and PM10 should be available and at least three out of the seven should be available.There is no theoretical upper value of AQI but it's rare to find values over 1000.The pre-defined buckets of AQI are as follows:



Dataset

The data for the concentration of pollutants are collected from the **Central Pollution Control Board (CPCB)** where data is publicly available for different locations across India. For this project

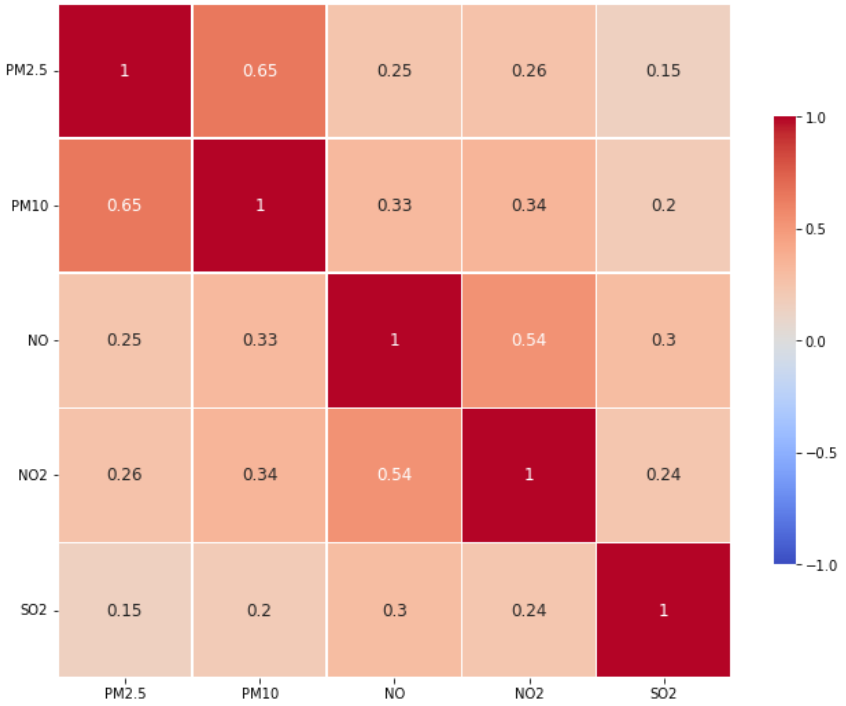
we chose Jodhpur City.All data used in this project are collected from sensors installed at CRT CAAQMS, Colectrate Campus Park, Paota High Court Road, **Jodhpur**.

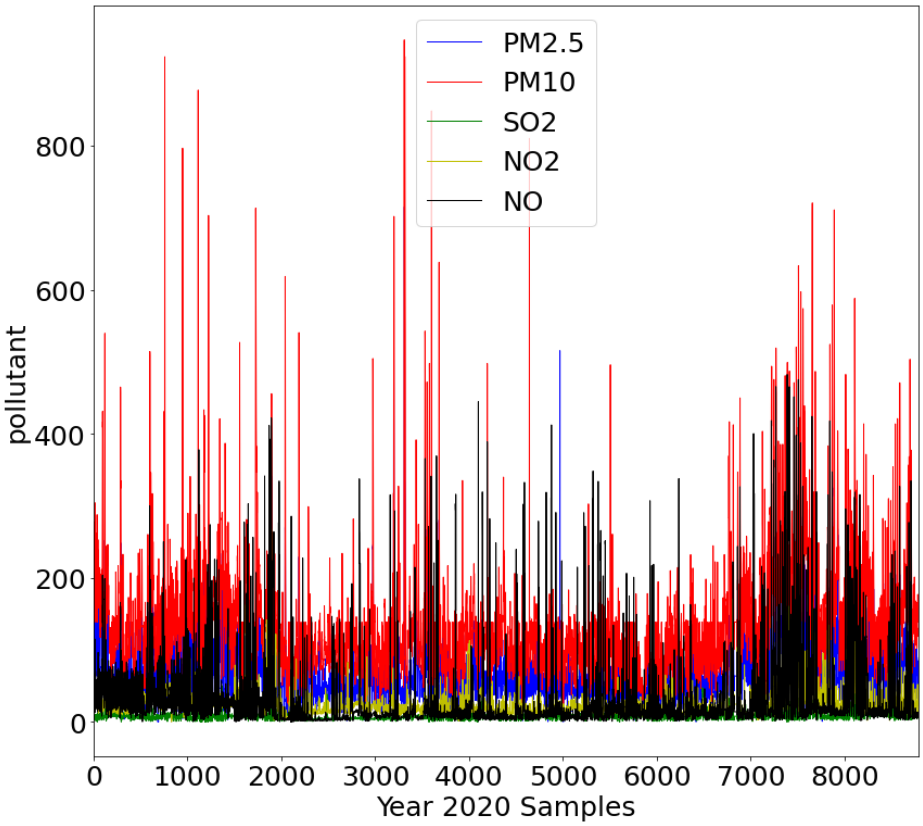
So, data for 3.25 years was selected for the purpose. The timeline of the data is from 1 January 2019 to 15 March 2022. We have around 28818 data points in our dataset.

Our dataset have **5 pollutants** and their statistics are defined below

	PM2.5	PM10	NO	NO2	SO2
count	28818.000000	28818.000000	28818.000000	28818.000000	28818.000000
mean	75.423606	158.772243	28.302659	29.973451	8.870322
std	55.468516	92.525423	47.856323	21.252670	5.421311
min	0.180000	0.300000	0.010000	0.040000	0.020000
25%	43.560000	103.600000	8.380000	16.800000	5.700000
50%	63.790000	138.650000	14.210000	24.740000	7.990000
75%	91.387500	187.777500	24.890000	36.300000	10.840000
max	999.990000	989.500000	488.700000	455.170000	171.330000

The correlations between all five pollutants can be described using following heatmap -





Deep Learning Approach -

Deep Learning is a class of machine learning algorithms that uses multiple layers to extract higher-level features from the raw input progressively. RNN is a popular deep learning architecture that is used to model sequential data. It contains cyclic connections where the outputs from previous time steps are fed as input to the current time step. In RNNs, errors are back propagated, and weights are updated using a technique called Backpropagation Through Time (BPTT).

There have been several approaches to predict air quality like ARIMA (Auto-Regressive Integrated Moving Average) and PCR (Principal Component Regression) statistical models. We will be using Special RNN network called LSTM to forecast 5 pollutants named (PM2.5, PM10, NO2, NO, SO2) and we will be further calculate AQI using formula explained above and then we will classify AQI in 6 categories and the will will evaluate our model based on RMSE of AQI value predicted as well as precision-recall F1-score between 6 classes of air quality.

LSTM -

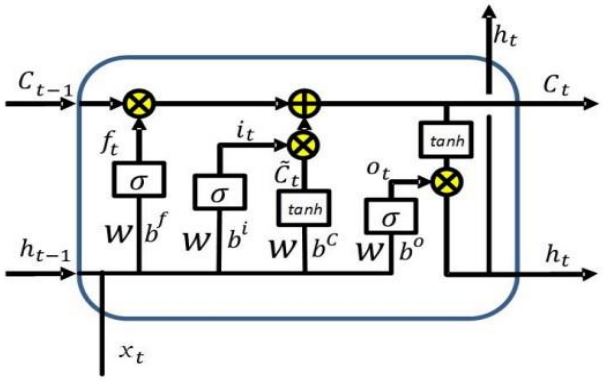
Long Short Term Memory Network is an advanced RNN, a sequential network, that allows information to persist. It is capable of handling the vanishing gradient problem faced by RNN. A recurrent neural network also known as RNN is used for persistent memory.

The core concepts of LSTM are the cell state, and its various gates. The cell state act as a transport highway that transfers relative information all the way down the sequence chain. You can think of it as the "memory" of the network. The cell state, in theory, can carry relevant information throughout the processing of the sequence. So even information from the earlier time steps can make its way to later time steps, reducing the effects of short-term memory. As the cell state goes on its journey, information gets added or removed to the cell state via gates. The gates are different neural networks that decide which information is allowed on the cell state. The gates can learn what information is relevant to keep or forget during training.

More specifically, the structure of a single layer LSTM network is depicted in Figure below. And the LSTM network updates itself at time step t as follows

$$
\begin{aligned}
f_t &= \sigma(W^f[h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W^i[h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= tanh(W^C[h_{t-1}, x_t] + b_C) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
o_t &= \sigma(W^o[h_{t-1}, x_t] + b_o) \\
h_t &= o_t * C_t,
\end{aligned}
$$

where ht is the hidden state; it , ft and ot refer to the input gate, forget gate and output gate, respectively; C˜ t and Ct refer to the input modulation gate and memory gate, respectively. {W f , Wi , WC, Wo} are the weights, and {b f , b i , b c , b o} are the biases for the corresponding gates; σ(·) and tanh(·) are sigmoid and hyperbolic tangent activation functions, respectively. The memory cell unit Ct contains two components, i.e., previous memory cell unit Ct−1 modulated by ft and C˜, which is modeled by the current input, and previous hidden state, modulated by the input gate it . The essence of sigmoidal operation for it normalizes themselves into the scope of [0,1]. Particularly, they could be deemed as knobs that LSTM learns to selectively forget its previous memory or consider its current input. In a similar way the output gate models the transfer from memory cells to hidden states.



(a) Structure of LSTM Network

Model Training -

Proposed LSTM model is developed using multiple python packages like keras, scikit-learn and tensorflow. The Min-Max Scaler of the scikit-learn library is used to normalize the data in the range between 0 and 1.

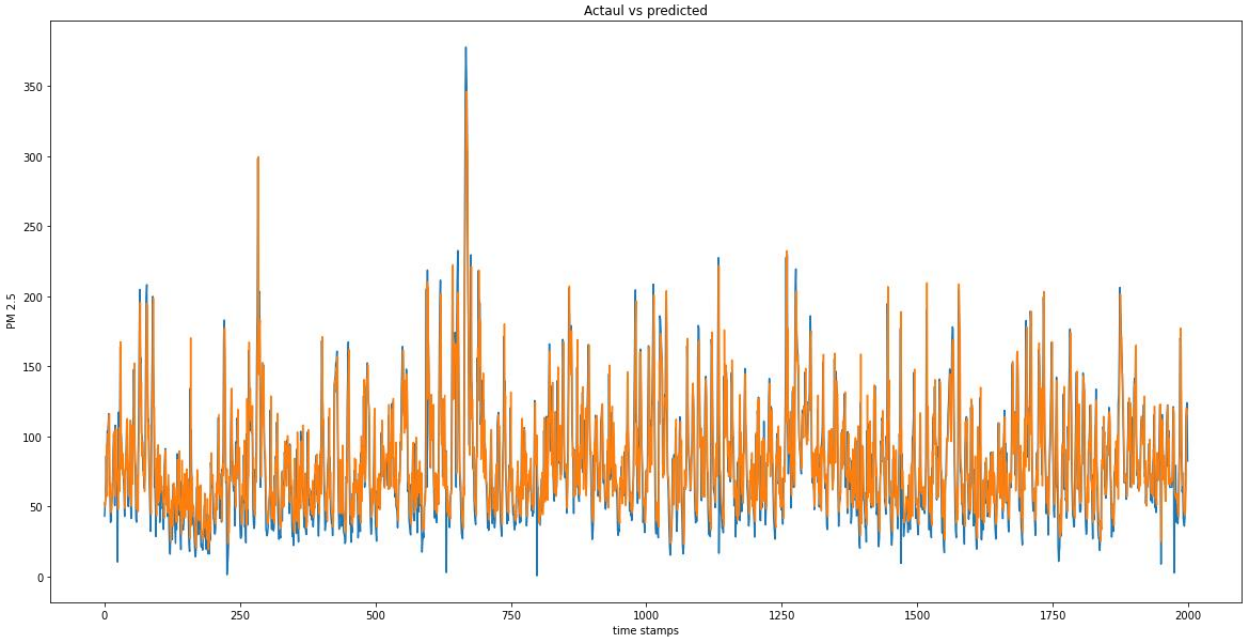
The architecture of the LSTM model depends on multiple parameters like number of epochs, batch size, number of LSTM layers, and the number of units in each LSTM layer. These parameters are adjusted such that there is a balance between underfitting and overfitting. Dropout Layers are also used to prevent models from overfitting. Dropout layers randomly drop units from the network during training and prevent units from co-adapting too much. The model is trained for 50 epochs. We used the relu activation function, adam optimiser and MSE error for fitting the model.

```
def lstm(df):
    n_steps = 4
    # split into samples
    X, y = split_sequence(df, n_steps)
    # reshape from [samples, timesteps] into [samples, timesteps, features]
    n_features = 1
    X = X.reshape((X.shape[0], X.shape[1], n_features))
    model = Sequential()
    model.add(LSTM(50, activation='relu', input_shape=(n_steps, n_features)))
    model.add(Dense(1))
    model.compile(optimizer='adam', loss='mse')
    # fit model
    model.fit(X, y, epochs=200, verbose=0)
    return model
```

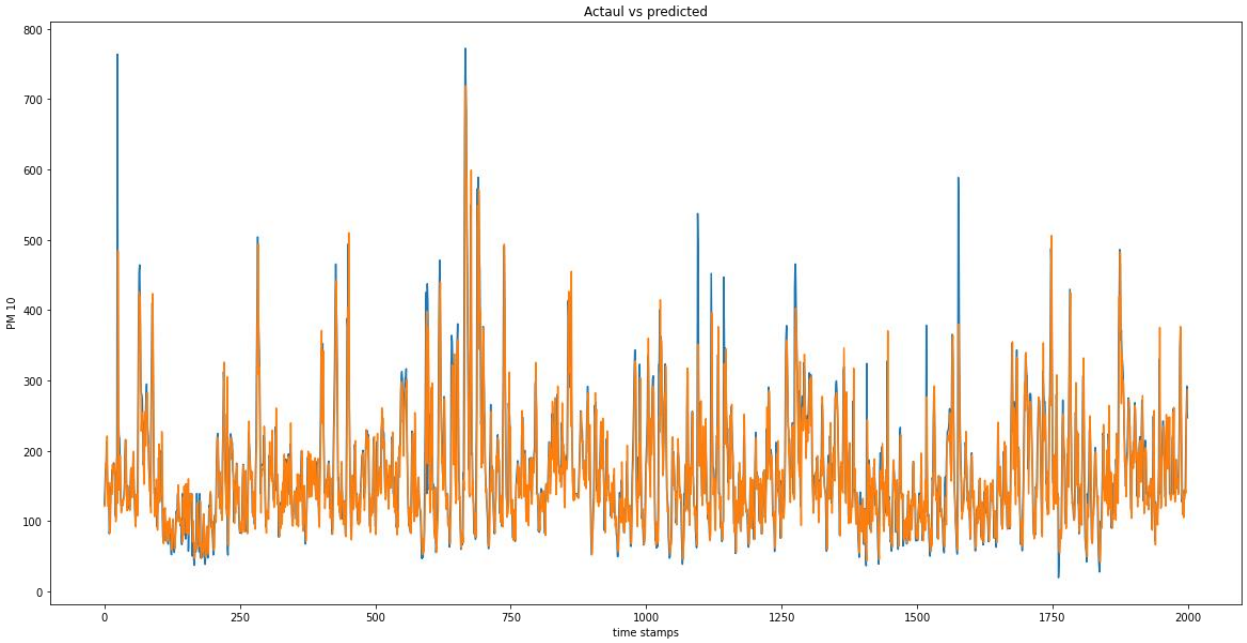
After training the LSTM based RNN models, each pollutants' concentration was predicted for the next hour using data from the last 4 hours.

We got the following distribution of each pollutant -

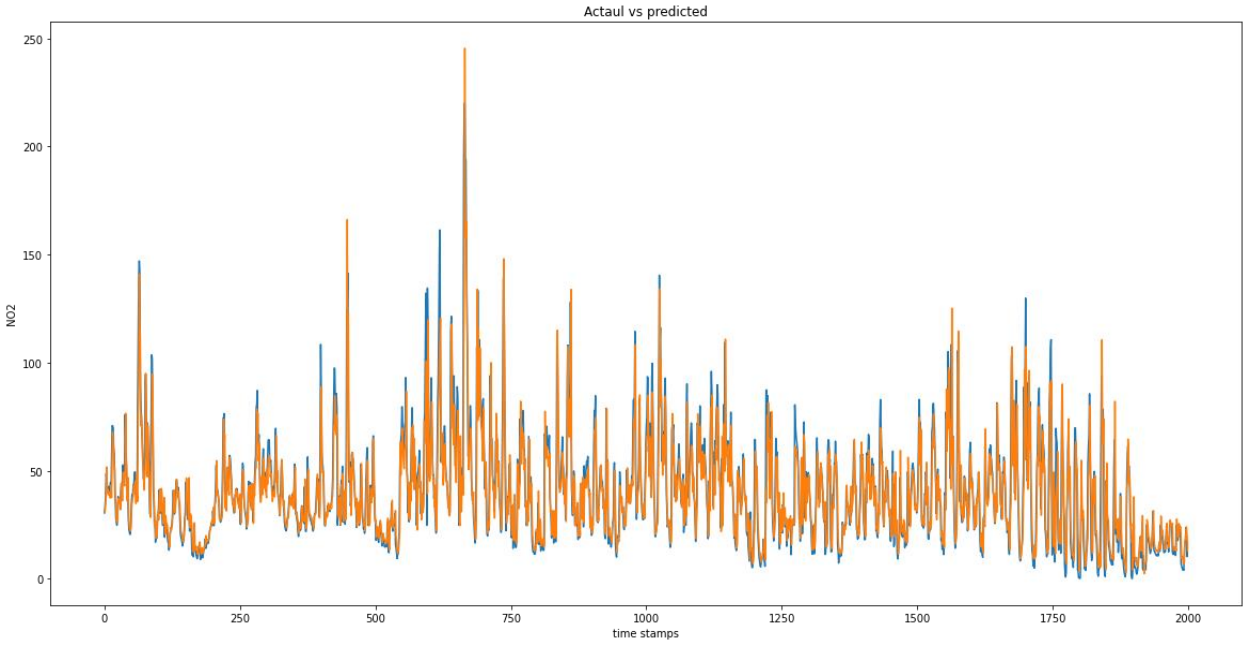
1. **PM2.5**



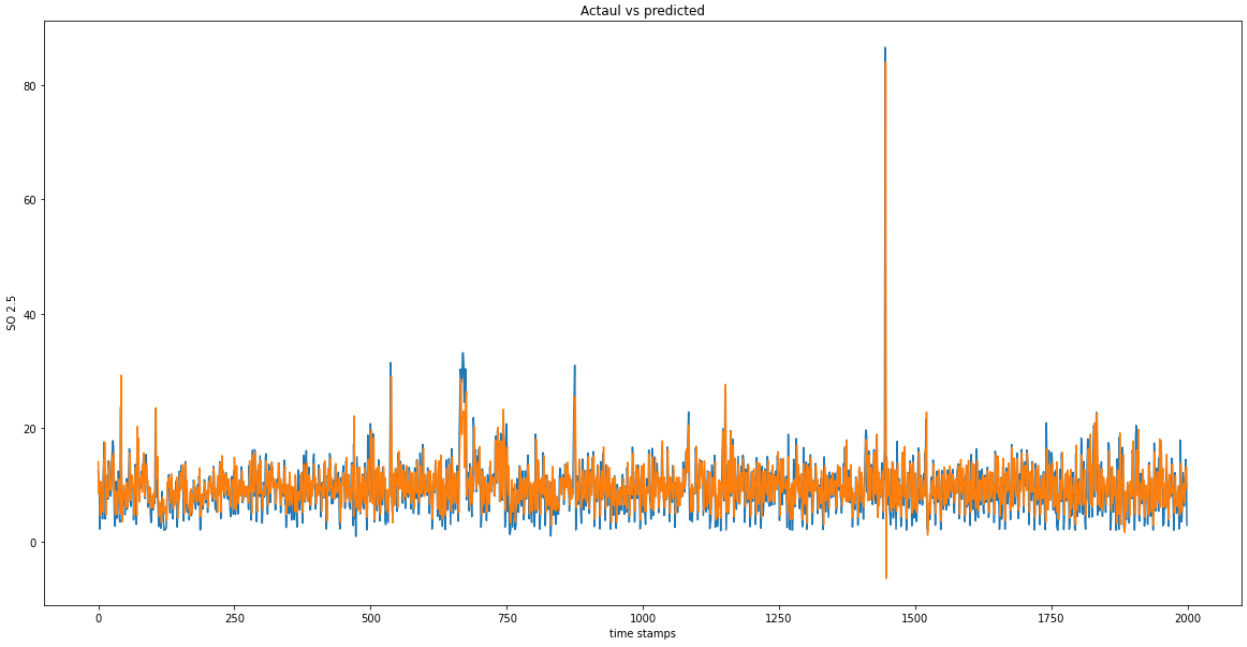
2. **PM10**



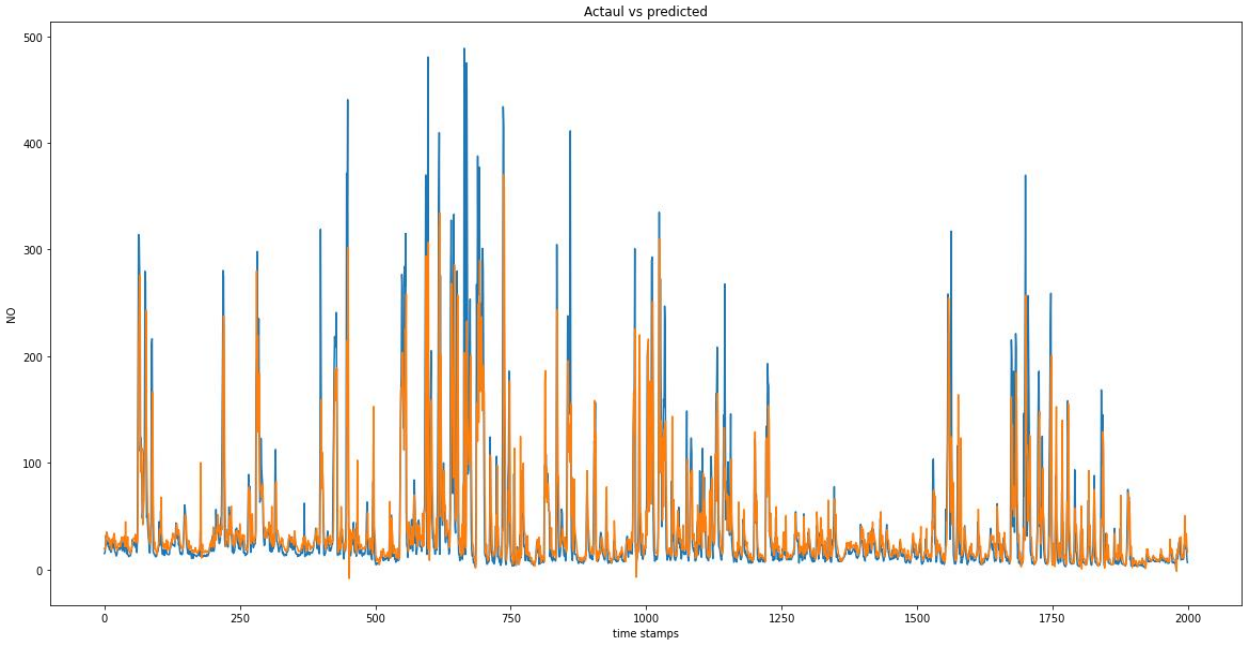
3. NO2


Actaul vs predicted
NO2
time stamps
0
50
100
150
200
250
0
250
500
750
1000
1250
1500
1750
2000

4. SO2


Actaul vs predicted
SO 2.5
time stamps
0
20
40
60
80
0
250
500
750
1000
1250
1500
1750
2000

5. NO



The RMSE value for our model is -
The RMSE is calculated between predicted AQI value and actual AQI values.

Pollutant	RMSE
PM2.5	25.43
PM10	27.12
NO2	19.48
NO	32.16
SO2	16.37

The evaluation metric for our model -

AQI category	Precision	Recall	F1- Score
Good	0.54	0.71	0.61

Satisfactory	0.89	0.84	0.86
Moderate	0.88	0.81	0.84
poor	0.91	0.86	0.87
Very poor	0.94	0.83	0.89
severe	0.91	0.76	0.82

Conclusion -

Establish an efficient forecasting model for AQI inJpdhpur.Proposed an RNN-LSTM model that predicts the hourly concentration of pollutants present in the air. The predicted concentrations are then used to calculate the AQI for a particular region's city. The present study is carried out on 3.5 years of hourly data from January 2019 to March 2022. The results show that deep learning-based techniques carry out more promisingly than conventional statistical methods. This work can be extended by predicting a higher number of future timesteps for all the eight pollutants considered in calculating AQI and also Temporal sequences of four meteorological parameters and pollutant levels can be fed as input to the LSTM model.

Codes and Dataset Used -

Code and dataset used in this project are uploaded on github at this link https://github.com/09sachin/AQI-forecating