Analysis and Prediction of Air Quality in India

Ms. Theres Bemila Jenet[1], Tushar Gada[2], Harshil Patel[3], Usashi Roy[4]

[1]Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, India
[2,3,4]Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India,

---

Abstract - *Growing air pollution in India over the years has resulted in causing crucial environmental and health-related hazards. There has been a need to inspect the air quality of the major commercial cities, which is contributing to this deterioration. In this paper, data from various sources has been taken to study the various industrial, vehicular, and household air contaminants and to observe the pattern over the years, which includes the effect of lockdown on air quality. A comparative analysis of various time-series algorithms has been done, and the best algorithm has been used to compute the future outcomes, in order to get an estimation of the forthcoming air quality. This research can then be used to take the measures necessary to get the scenario under control.*

Key Words*: *Air Quality Index, Data Pre-Processing, Analysis of Air Quality, Python

1. INTRODUCTION

As a developing country, India has seen a rise in the number of modernizing cities, which has increased technological and industrial growth. This rapid development has led to the manufacturing of many vehicles and devices that contribute to the emission of toxic gases into the atmosphere. Over emission of these gases has resulted in serious issues such as smog, acid rain, and so on. Due to deforestation in urban localities, the air inhaled has become a serious concern. The air quality of some cities has reached a certain critical point set by the government. An effective study of air quality is required to instill public awareness.

This research includes data which was taken from the Central Pollution Control Board (CPCB). The CPCB has developed the Air Quality Index (AQI) parameter to inspect the daily air contamination levels of urban areas. The AQI gives us the actual qualitative form of the air around us so we can associate it with the various impacts on our health. The pollutants such as Particulate Matter (PM2.5 and PM10), O3, NO2, NO, NOx, NH3, CO, and SO2 of the metropolitan cities have been taken into consideration. Based on the concentration of these gases, the AQI level of the individual cities has been calculated. Furthermore, exploratory data analytics of these gases have been performed to get a better view of the prime pollutants affecting the air quality in the metropolitan cities from 2018 to 2021. The dataset being used can be classified as time-series data. To perform forecasting, we found the most accurate algorithm by applying various algorithms to the dataset.

1.1 Problem Statement

To get knowledge about air quality, it is necessary to back-track the major pollution-causing pollutants and the locations affected seriously by the pollutant across India. The analysis of underlying principles and patterns will give us an insight into future air quality measures. Moreover, as there has been a drastic change in the Air Quality and the concentration of other emissions of gases lately because of the pandemic situation of the year 2020, it becomes necessary to inspect the overall changes over time.

1.2 Scope of the Project

The Analysis and Prediction of Air Quality in India is a data science research-based project which includes data analysis and forecasting using historical data. The historical data includes data taken from a government-authorized website i.e. The Central Pollution Control Board which we use for our predictions. Based on the data collected from the official site, the Dashboard will form multiple geographical graphs, Heat maps, and other visualizations. The project would also provide an estimation of the Air Quality of the major cities in India for the next consecutive year.

1.3 Purpose

Most of the predictions have been done on the current data and not forecasted. As India is a developing nation, with an increasing number of factories and industries, Air Quality Index is an essential factor to be considered. As there is a huge

difference between the AQI of our country before and after the lockdown, analyzing this data gives us a better overview of the scenario with appropriate visualizations.

Fig -1: Flowchart of the Project

2. DATA PREPROCESSING

After gaining some domain knowledge, we analyzed several factors by which the air quality index is calculated. The raw data taken from the official government website was first converted into excel format, which included the city names, dates, and the gases.

The data collected was sparse as it consisted of null values and outliers, which were treated using linear interpolation. After correcting the data, we calculated the AQI using the general formula.

Using the AQI, we made another column that included the severity status of the air quality level as provided by the Central Pollution Control Board and State Pollution Control Board under the Swachh Bharat Abhiyan.

Table -1: AQI Category and Pollutants

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
Moderate (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
Severe (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Hazardous (401+)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

For analysis, i.e., for geographical distribution, other attributes like the latitude and longitude of the cities have also been added to the dataset.

Table -2: Dataset Content

Attributes	Description	Data Type	Duration
City	26 cities of India	Nominal	-
Date	2015 to 2020	Datetime	-
$PM_{2.5}$	Particulate Matter 2.5-micrometer in ug / m3	Numeric	24-hr avg
PM_{10}	Particulate Matter 10-micrometer in ug / m3	Numeric	24-hr avg
NO	Nitric Oxide inS ug / m3	Numeric	24-hr avg
NO_2	Nitric Dioxide in ug / m3	Numeric	24-hr avg
NO_x	Any Nitric x-oxide in ug/m3	Numeric	24-hr avg
NH_3	Ammonia in ug / m3	Numeric	24-hr avg
CO	Carbon Monoxide in mg / m3	Numeric	8-hr max
SO_2	Sulphur Dioxide in ug/ m3	Numeric	24-hr avg
O_3	Ozone ug / m3	Numeric	8-hr max
AQI	Air Quality Index	Numeric	-
AQI Bucket	Air Quality Index Status	Nominal	-

3. ANALYSIS

The daily results of the index are used to convey to the public an estimate of the level of air pollution. The analysis of the historical data is done to get a better view of the prime pollutants affecting the air quality, the locations that were majorly affected, and the variation of pollution levels over the time from 2018 to 2021. Moreover, an analysis has been performed on the effect of lockdown on air quality, which compares the scenarios before and after the year 2020, when there was a worldwide shutdown. Methods of visualization used for this purpose were pyplot, plotly (plotly express and graph objects), and geoplotting (using folium and choropleth maps).

4. FORECASTING USING ALGORITHMS:

Based on the dataset, which included data over time, we used time-series algorithms to predict the AQI (up to August 2022). Moreover, we used the historical data to predict the scenario where lockdown would not exist by dividing the data into two parts where the training data has been taken before March 2020 and the testing data from March 2020 to August 2021, and the testing part has been predicted and compared. The algorithms used are as follows:

4.1 ARIMA (Autoregressive Integrated Moving Average)

Just like ETS, ARIMA / SARIMAX is part of the old yet very good Forecasting Methods for Time Series. It also provides an excellent base and is easy to execute using one line R or Python. It is also embedded in the Alteryx Desktop. To use Python for ETS and ARIMA models, you can use the statsmodel package.

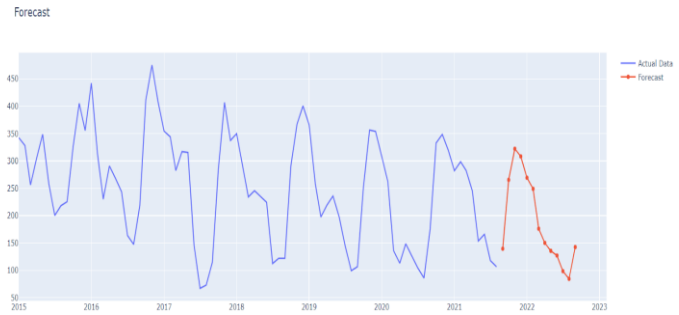

Chart -1: ARIMA Forecasting of the year 2022 for Delhi

4.2 Prophet by Facebook

Facebook updated its Time Series algorithm for 2017: Prophet. It is very easy to use with a few Python or R lines and provides an easy-to-interpret prediction, the algorithm is not too complicated. Compared to ETS it is able to cope with changing trend patterns and is better designed for high-frequency data (daily or more). Users can also incorporate some business acumen by placing the floor and ceiling in the forecast, and the Prophet uses this information to reduce the inclination accordingly if necessary.

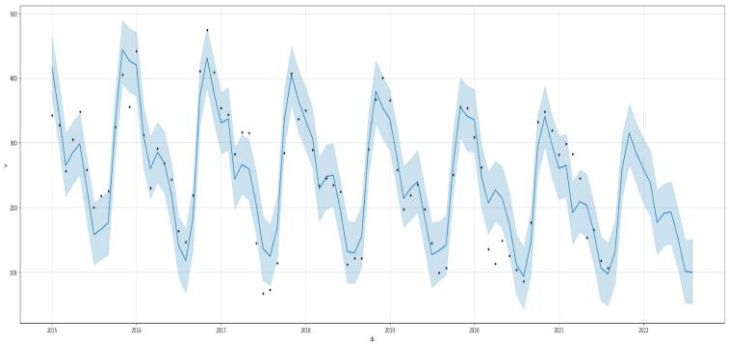

Chart -2: Prophet Forecasting of the year 2022 for Delhi

4.3 LSTM (Long Short-Term Memory Algorithm)

Deep Learning offers exciting ways to predict the Time series. Among them, Recurrent Neural Networks (RNN) and LSTM cells (Long-Short Term Memory) are popular and can be used with a few lines of code using Keras as an example.

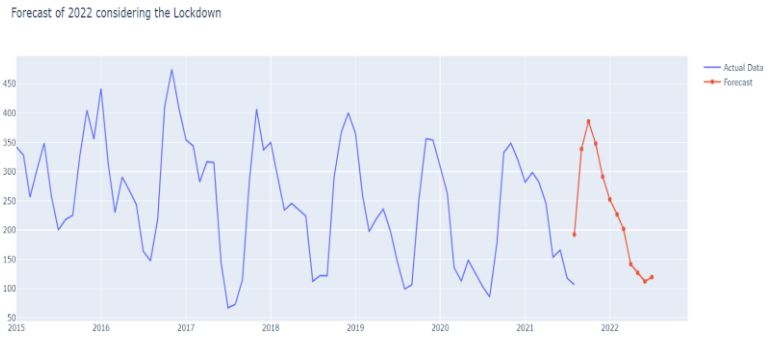

Chart -3: LSTM Forecasting of the year 2022 for Delhi

4.4 ETS (Exponential Smoothing)

Exponential Smoothing prediction or ETS algorithm is one of the simplest and fastest to predict time series very accurately. It can manage trends and seasons and is easy to interpret. It can be used with a single line of code in R or Python and comes embedded in tools like Alteryx. Its version of AAA is also embedded in Excel as a function of FORECAST.ETS and PowerBI to predict Line charts.

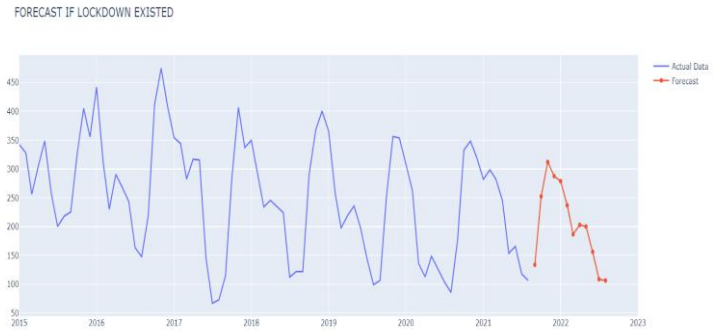

Chart -4: ETS Forecasting of the year 2022 for Delhi

5. COMPARATIVE ANALYSIS

The output of the most compatible algorithm concerning the dataset is represented in the dash application. After implementing all the algorithms in all the cities it was noticed that the rolling forecast of the ARIMA algorithm gave us the best output. The Exponential Smoothing Algorithm gave us a comparable output followed by the Prophet algorithm. On the other hand, for the RNN/LSTM algorithm, since the entries in the dataset were not enough for the algorithm, it did not seem to give the expected accuracy. Parameters like RMSE, MAPE, MSE, ME, and MPE were considered.

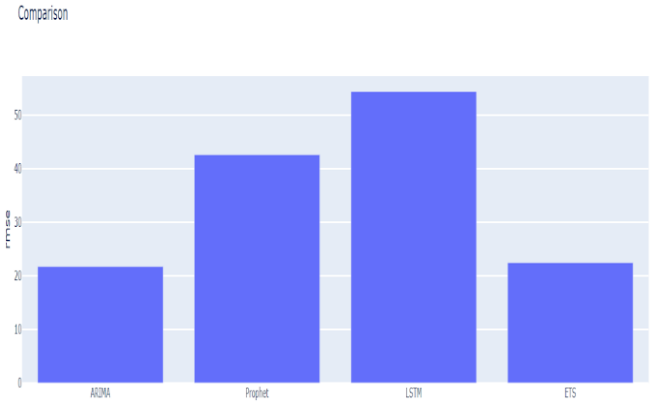

Chart -5: Comparative analysis (RMSE) of all the Algorithms on Delhi dataset

6. DASH (USER INTERFACE)

As we were using several visualizations and tables in our project, we found Dash to be most suitable for our project. Dash is an open-source framework for creating data view links. The use of dashboards in organizations and the development of dashboard applications have continued to increase these days. We can portray the huge amount of data and graphs in a minimal and understandable format using the dashboards.

The dashboard includes citywide analysis and algorithm implementation, which provides the forecast and the comparative analysis of each algorithm. The data analysis also provides us with geographical visualizations.

Fig -2: Page for Amritsar on Dash Application

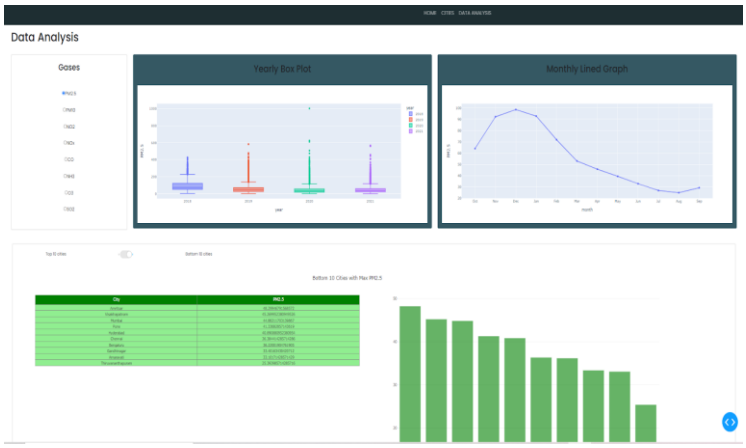

Fig -3: Page for Data Analysis on Dash Application

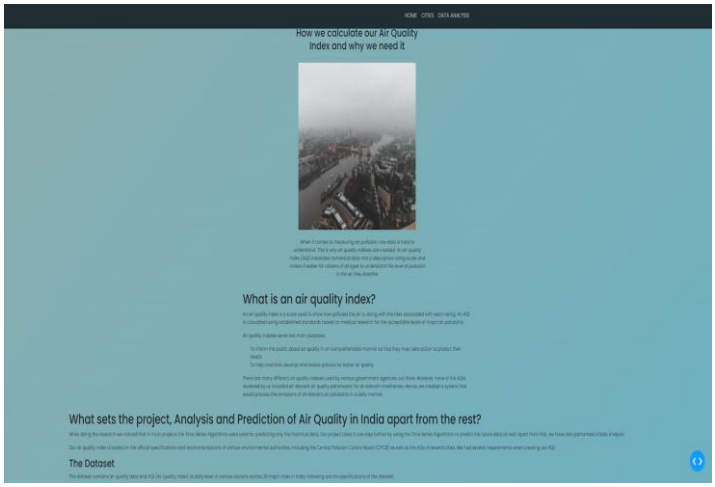

Fig -4: Home Page from Dash Application

7. CONCLUSION & FUTURE SCOPE

ARIMA, LSTM, Prophet, and ETS are very good Forecasting Methods for Time Series which provide a very good baseline that is implemented using python. The Comparative analysis of the algorithms provided us with the most efficient algorithm, (ARIMA Algorithm) for forecasting. Along with the Air quality, the intensity of the pollutants can also be predicted. The temperature changes in our country and the health effects can also be analyzed.

8. SUMMARY

Air quality indicator (AQI) is a dimensionless indicator that statistically describes the state of air quality. By predicting air quality indicators, we can reverse pollution-causing pollution and the worst-affected area across India. With this forecasting model, various knowledge about the data is extracted using various techniques to obtain heavily affected regions in a particular region (cluster). This gives more information and knowledge about the cause and seniority of the pollutants.

REFERENCES

[1] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches", International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018
Link: http://www.ijesd.org/vol9/1066-C0049.pdf

[2] Zhou Kang, Zhiyi Qu, "Application of BP Neural Network Optimized by Genetic Simulated Annealing Algorithm to Prediction of Air Quality Index in Lanzhou", 2017 2nd IEEE International Conference on Computational Intelligence and Applications
Link:https://twin.sci-hub.se/6684/7f459e29300678fc97e95c3464cae177/kang2017.pdf

[3] Yuchao Zhou;Suparna De;Gideon Ewa;Charith Perera;Klaus Moessner, "Data-Driven Air Quality Characterization for Urban Environments: A Case Study", Year: 2018 |Volume: 6 | Journal Article |
Link: https://ieeexplore.ieee.org/document/8555540

[4] Venkat Rao, Uhasri, Pavan Kalyan Srikanth, Hari Kiran Reddy, "Air Quality Prediction Of Data Log By Machine Learning", 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)
Link: https://sci-hub.se/10.1109/ICACCS48705.2020.9074431

[5] A. Gnana Soundari, J. Gnana Jeslin M.E, Akshaya A.C, "INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue),
Link: https://www.ripublication.com/ijaerspl2019/ijaerv14n11spl_34.pdf

[6] Kalash Agarwal, Yatender Singh, Jasmendra Singh, Abhishek Goyal, "Air Pollution Prediction Using Machine Learning",International Research Journal of Engineering and Technology (IRJET),Volume: 07 Issue: 07 | July 2020
Link: https://www.irjet.net/archives/V7/i7/IRJET-V7I7293.pdf

[7] Radhika M.Patil, Dr. H. T. Dinde, Sonali. K. Powar, "A Literature Review on Prediction of Air Quality Index and Forecasting Ambient Air Pollutants using Machine Learning Algorithms", International Journal of Innovative Science and Research Technology, Volume 5, Issue 8, August – 2020
Link: https://www.ijisrt.com/assets/upload/files/IJISRT20AUG683.pdf

[8] Jason Brownlee, "Deep Learning for Time Series Forecasting,Predict the Future with MLPs, CNNs and LSTMs in Python", Originally published: 30 August 2018
Link: https://www.google.co.in/books/edition/_/o5qnDwAAQBAJ?hl=en&gbpv=1

[9] SABA AMEER, MUNAM ALI SHAH , ABID KHAN , HOUBING SONG , CARSTEN MAPLE, SAIF UL ISLAM, AND MUHAMMAD NABEEL ASGHAR, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities", SPECIAL SECTION ON URBAN COMPUTING AND INTELLIGENCE, date of publication June 26, 2019.
Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8746201

[10] https://github.com/krishnaik06/AQI-Project

[11] Dataset: https://www.kaggle.com/rohanrao/air-quality-data-in-india

[12] AQI Calculation: https://www.kaggle.com/rohanrao/calculating-aqi-air-quality-index