International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper | Volume-9, Issue-9, 2022 Regular Edition | E-ISSN: 2349-7084

Forecasting Air Pollution Concentrations and Binning Air Quality Index Values to Encourage Green Vehicles for Sustainability: A Data Science Approach

[1]Bushitha Reddy Baddam,[2] D.Shivani, [3]Kambalapally Sriya Reddy, [4]T.Sriya, [5]G.Deepika

[1,2,3,4]B.Tech , IV[th] Year , Department of Computer Science and Engineering, CVR College of Engineering, Vastunagar, Mangalpally, Ibrahimpatnam, T.S., India – 501510.

[5]Assistant professor, Department of Computer Science and Engineering, CVR College of Engineering, Vastunagar, Mangalpally, Ibrahimpatnam, T.S., India – 501510

Email ID: [1] bushithareddy1234@gmail.com , [2]shivaniidonuru139@gmail.com, [3]reddysriyak12@gmail.com, [4]shriyatreddy@gmail.com, [4]deepika.gundala@gmail.com

Corresponding author : ***G.Deepika***

Available online at: http://www.ijcert.org

Received: 15/10/2022, Revised: 27/11/2022, Accepted: 12/12/2022, Published: 22/12/2022

Abstract:. People don't get as much clean air as they used to because of pollution. Contaminated air is harmful since it can lead to respiratory and cardiovascular problems. The data science process is used to deal with this issue. This method assists in the systematic analysis of air pollutants that influence Air Quality Index (AQI) values. The primary objective of this research is to utilize data science in order to make long-term AQI predictions for the city of Hyderabad. To accomplish this goal, pre-COVID-19 and post-COVID-19 AQI data are combined into a dataset. The data science methodology is applied to solve this issue. Through this method, air pollutants that have an impact on the Air Quality Index (AQI) can be analyzed in a methodical fashion. The primary objective of this research is to utilize data science in order to forecast future AQI values for the city of Hyderabad. This is done by assembling a database of AQI readings from both before and after the onset of COVID-19. First, the data is cleaned, and then exploratory data analysis (EDA) is performed to better understand when and why varying air pollutants have changed over time. In addition to training the sophisticated forecasting model, the seasonal auto-regressive integrated moving average with exogenous factors (SARIMAX) is also trained with these trend and seasonality components. This model forecasts the amount of air pollution in the following three years. The severity of air pollution in a city is evaluated by aggregating the estimated AQI values across the AQI categories. Based on these results and how they can be interpreted, we want to encourage people to purchase environmentally friendly vehicles so that we can live in a sustainable manner.

Keywords: Air Quality Index, pre-COVID-19, exploratory data analysis, Seasonal Auto-Regressive Integrated Moving Average with eXogenous Factors, data science.

---

1. Introduction

Air pollution is considered to occur whenever harmful or excessive quantities of defined substances, such as gases, particulates, and biological molecules, are introduced into the atmosphere. The effects of these high emissions are clear: they make people and other living things sick, kill them, and hurt crops. More than 90% of India's population lives in areas where air quality is below World Health Organization standards, with coal-fired power plants, factories, and vehicles among the major sources of pollution. The most common primary pollutants are sulfur dioxide (SO2), particulate matter (PM), nitrogen dioxide (NOx), and carbon monoxide (CO). The two

primary sources of PM2.5 pollution in the city are automobiles and industries. Vehicles contribute one-third of the pollution. Cars are a big cause of air pollution because they make a lot of nitrogen oxides, carbon monoxide, and small particles. 80–90% of cars' environmental impact comes from fuel consumption and emissions of air pollution and greenhouse gases. A time series forecasting algorithm is used to predict these values using AQI. AQI, which stands for "Air Quality Index," is a measure of pollutants in the air. It shows how bad the air quality is by saying it is moderate, satisfactory, good, bad, very bad, or severe.

A higher value of the AQI indicates more air pollution. An AQI between 0 and 50 is considered good, 51 and 100 satisfactory, 101 and 200 moderate, 201 and 300 poor, 301 and 400 very poor, and 401 and 500 severe. AQI helps analyze changes in air quality (improvement or degradation) and informs the public about environmental conditions. It is especially helpful for people with illnesses that are made worse or caused by dirty air.

Delhi, the capital city of India, has the highest AQI value in India and the world. Because of the high AQI, Delhi has given students a week off. The main contributors to air pollution are vehicles, which emit various pollutants.

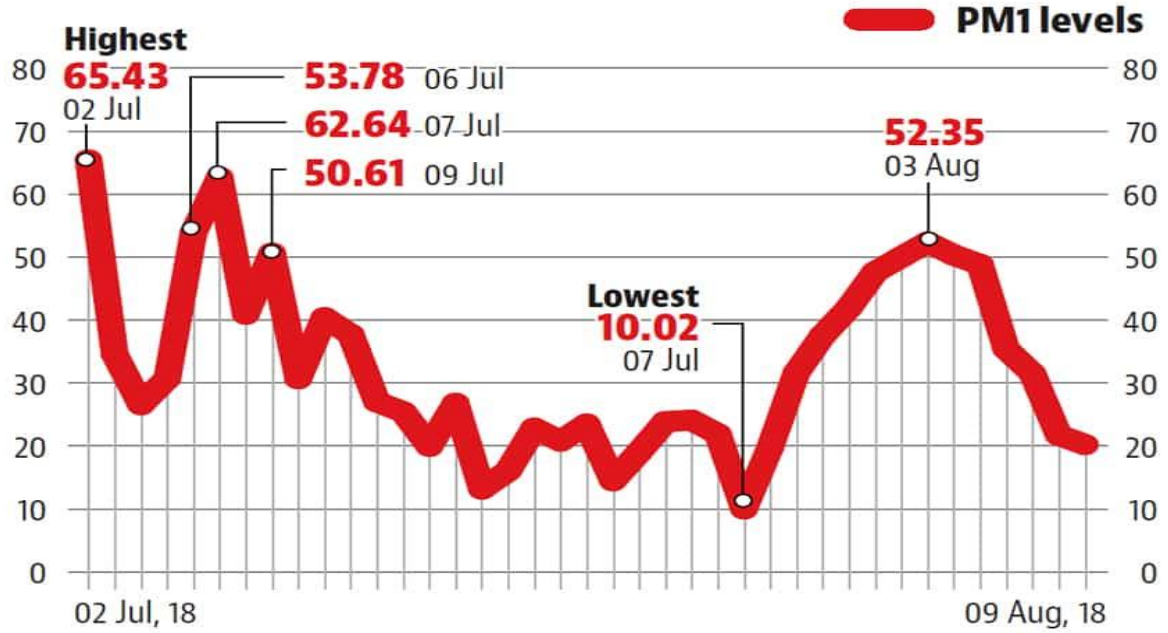

Figure 1: Represents Average AQI in Delhi

Air quality indices have been created in different countries for the measurement of air quality. These indices measure the air quality in the country and indicate whether the amount of nitrogen dioxide, carbon monoxide and sulfur dioxide in the air exceeds the criteria set by the World Health Organization or not.

India uses the National Air Quality Index (AQI), Canada uses the Air Quality Health Index, Singapore uses the Pollutant Standards Index and Malaysia uses the Air Pollution Index.

There are many cities including Beijing, Paris where 'pollution emergency' is declared. However, India also declared the same in November 2019.

Particulate Matter (PM10):

Particles are defined by their diameter for air quality regulatory purposes. Those with a diameter of 10 microns or less (PM10) are inhalable into the lungs and can induce adverse health effects.

Particulate Matter (PM2.5):

As PM 2.5 simply refers to the size of particulate matter, an objective measurement can be taken. PM 2.5 AQI will indicate the number of these small particles in the air. This allows for a variety of insights regarding air quality.

Nitrogen Dioxide (NO2):

Nitrogen dioxide, or NO2, is a gaseous air pollutant composed of nitrogen and oxygen and is one of a group of related gases called nitrogen oxides, or NOx. NO2 forms when fossil fuels such as coal, oil, gas or diesel are burned at high temperatures. NO2 and other nitrogen oxides in the outdoor air contribute to particle pollution and to the chemical reactions that make ozone. It is one of six widespread air pollutants that have national air quality standards to limit them in the outdoor air.

Sulphur Dioxide (SO2):

The largest sources of sulfur dioxide emissions are electricity generation, industrial boilers, and other industrial processes such as petroleum refining and metal processing. Diesel engines are another major source, including old buses and trucks, locomotives, ships, and off-road diesel equipment.

 https://doi.org/10.22362/ijcert/2022/v9/i11/v9i1103

Carbon monoxide (CO):

It is an odorless, colorless, and tasteless but dangerous gas. Carbon monoxide is produced when fuels are burned such as gasoline, natural gas, oil, kerosene, wood or charcoal. Breathing CO reduces the blood's ability to carry oxygen. It can reach dangerous levels indoors or outdoors.

Ozone:

Ozone (O3) is a gas molecule composed of three oxygen atoms. Often called "smog," ozone is harmful to breathe. Ozone aggressively attacks lung tissue by reacting chemically with it. When ozone is present, there are other harmful pollutants created by the same processes that make ozone.

The ozone layer found high in the upper atmosphere (the stratosphere) shields us from much of the sun's ultraviolet radiation. However, ozone air pollution at ground level where we can breathe it (in the troposphere) causes serious health problems.

Ammonia (NH3):

NH3 is a precursor of PM2.5 which deteriorates urban air quality, affects human health and impacts the global radiation budget. Since vehicles are important sources of NH3 in urban areas

Lead (PB):

Lead in the air is emitted as aerosol predominately by burning of solid fuel (i.e. coal and biomass) and roasting of pyrite minerals in this region.

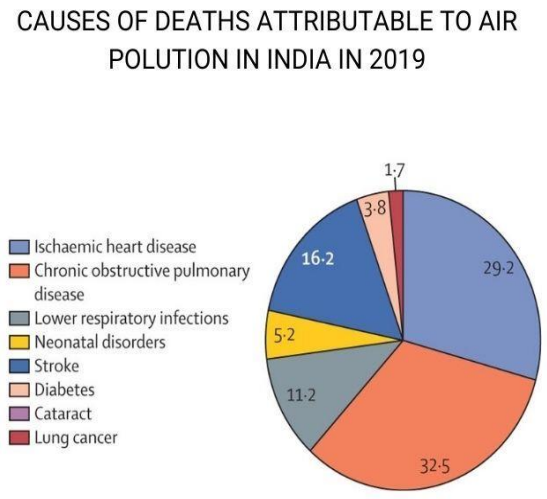

Figure 2: Causes of death due to air pollution pie chart

The impact of High AQI is hazardous to the citizens of the country, figure 2 represents various causes of deaths to air pollution in which deaths due to cardiovascular diseases and chronic obstructive pulmonary diseases constitutes two third of the entire pie chart. It is evident that High AQI has major impact on public health and in future it may be catastrophic. Without proper measures it may soar in future, so proper measures is needed from the government and international organisations to tackle it.

The remaining paper is organized as follows: Section 2 represents a literature review; Section 3 presents a proposed model; Section 4 presents a result analysis; and Section 5 presents concussion.

2. Literature Review

This article [1] focuses mostly on the short-term effects of air pollution on human health. People thought that reports of respiratory, cardiovascular, and cerebrovascular symptoms were linked to keywords about diseases linked to air pollution. The authors used the Baidu index instead of hospital outpatient data to model how air pollution affects people's health. They did this because they thought it would represent people searching for respiratory, brain, and heart problems symptoms. Respiratory, cardiovascular, and cerebrovascular symptoms were linked to keywords about diseases linked to air pollution. The authors used the Baidu index instead of hospital outpatient data to model how air pollution affects people's health. They did this because they thought it would represent people searching for respiratory, brain, and heart problems symptoms. We also compared the Ministry of Environmental Protection's (MEP's) existing air quality index (AQI) for China, which is based on Baidu (China's most popular search engine), to an index based on the relative risk (RR) that is intended to better identify the health dangers of air pollution (specifically respiratory, cerebral, and cardiovascular diseases). Higher search volumes were seen for terms like "bronchitis," "asthma," "lung cancer," "pneumonia," "rheumatism," and "cough," except for "asthma," which didn't show any clear cycles or search indices related to the heart and brain. Searches for terms about cardiovascular illness were the most popular.

The author of this study suggests [2]. Concerns about one's health make urban travel even more dangerous. The authors of this study talked about how important air quality index measurements are for human health. They said that when the index is particularly low, these readings can be used to help people choose safer ways to travel. We employed the Dijkstra algorithm to find the safest path from point A to point B. Dijkstra's algorithm has been implemented because it can be used to find the safest path from the starting point to the ending point. Dijkstra's algorithm uses this number to determine which node has the worst air quality and which paths between nodes have the least pollution overall.

This study [3] presents evidence supporting the concept that even short-term exposure to ozone, nitrogen dioxide, sulphur dioxide, particulate matter (PM2.5), and total respiratory aerosol particles (TRAP) increases the incidence of asthma exacerbations. More and more evidence shows that asthma is linked to long-term exposure to air pollution, especially TRAP and its

 https://doi.org/10.22362/ijcert/2022/v9/i11/v9i1103

replacement, nitrogen dioxide. There needs to be a lot of research to prove that oxidative stress and immunological dysregulation are factors in asthma attacks caused by pollution. Poor children with asthma in low-income communities are more at risk. Suppose governments in developed and developing countries work together to wean their economies from burning fossil fuels for transportation and energy production. In that case, the air quality could improve, and the number of new asthma cases could decrease. To mitigate climate change, this approach is also essential.

The author of this work [4] uses the cost-of-illness method to determine the economic effects of air pollution in India. He or she does this by figuring out the cost of lost productivity due to early death and illness caused by pollution in each state of the country. Also, they figured out the economic loss as a percentage of the state's GDP. This ranged from 2% to 12% across the states, and it was highest in Uttar Pradesh, Bihar, Rajasthan, Madhya Pradesh, and Chhattisgarh, which had the lowest GDP per person. It is estimated that in 2019, the cost to the economy per person in Delhi due to air pollution was 54 times higher than in Haryana. They thought that India's goal of having a $5 trillion GDP by 2024 might not be possible because air pollution causes many deaths and illnesses and has a big negative effect on the economy when production goes down. Using strategies specific to each state in India to reduce air pollution would have huge economic and health benefits for the country.

3. Proposed Model

The significance of high-quality air is discussed. The day-by-day AQI dataset will be used for this purpose; it contains information on the typical pollution and AQI levels in around 26 cities in India between 2015 and 2020[5]. We have been doing similar work on how the COVID-19 quarantine has affected air quality in Hyderabad. Similarly, we may anticipate India's AQI in 2021, 2022, and 2023 by fitting a SARIMA model with computed orders to a time series of data. If so, we'll know for sure that air quality has improved most dramatically in Hyderabad City. Datasets are used to teach machine learning algorithms like regression and SARIMAX[6]. To better predict the data for each pollutant, we calculated the root mean square and got a result of 2.49, which is encouraging.

System Architecture

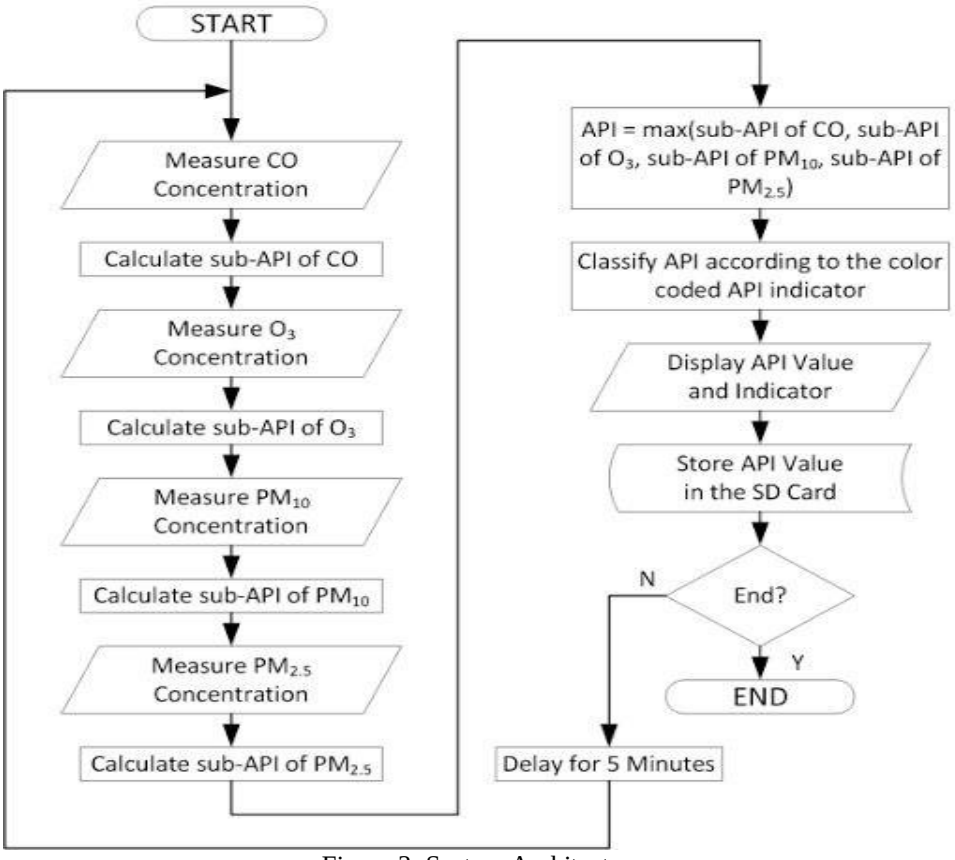

Figure 3: System Architecture

- Seven different variables are used to determine the AQI, as shown in Figure 3: PM2.5, PM10, SO2, NOx, NH3, CO, and O3.
- If at least 16 measurements have been taken over the course of the past week, an average will be

 https://doi.org/10.22362/ijcert/2022/v9/i11/v9i1103

calculated for PM2.5, PM10, SO2, NOx, and NH3.

- Sub-indices are derived from the original measures by classifying them into categories.
- When necessary data points are missing or there is a dearth of measurements, it can be difficult, if not impossible, to draw reliable conclusions[7].
- As long as one of PM2.5 and PM10 is measured and at least three of the seven are measured, the maximum of these two will be used to calculate the final AQI.
- The AQI readings must then be sorted using the color-coded scale.
- Displaying and saving the API value to an SD card would be the final step.

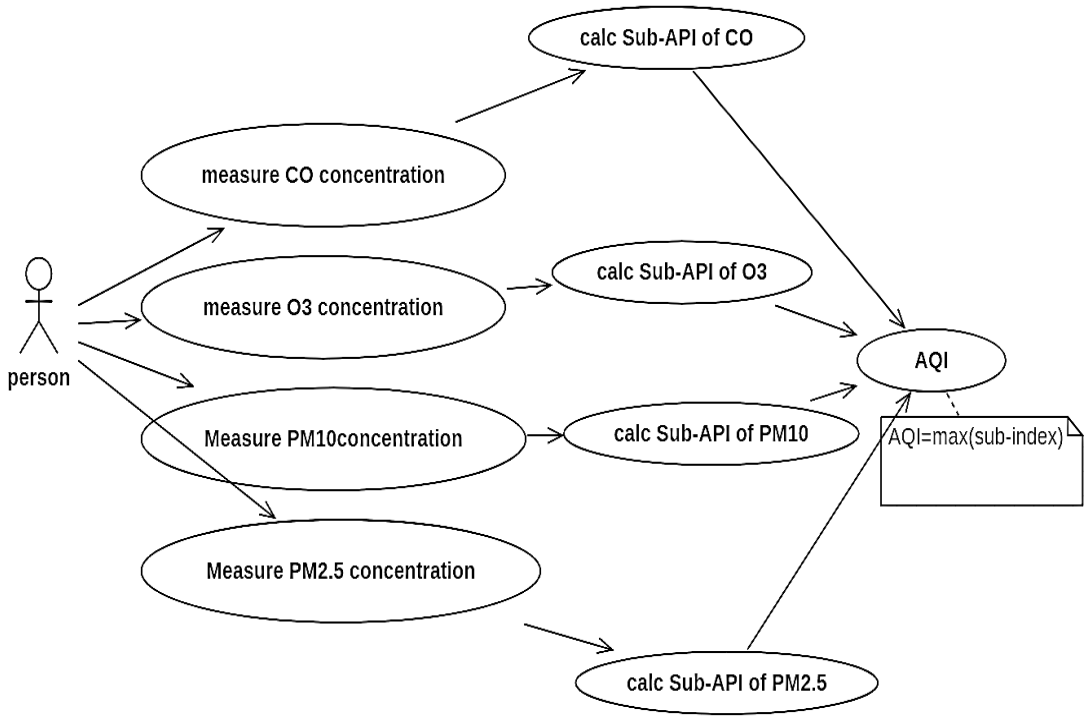

Figure 4: Usecase diagram

3.1 SARIMAX model

People have said that time-series forecasting with the SARIMAX model is the best way to make predictions and group the values.

SARIMAX, or the Seasonal Auto-Regressive Integrated Moving Average with Exogenous Factors, is a more recent refinement of the ARIMA model. SARIMAX incorporates seasonal effects and other external factors in addition to the autoregressive and moving average components, whereas ARIMA relies on an autoregressive integrated moving average. Since SARIMA and Auto ARIMA are also seasonal models, we can say that SARIMAX is the seasonal equivalent of the two[8].

An ARIMA (autoregressive integrated moving average) model is a refinement of an ARMA model used in statistics and time series analysis. A moving average and an autoregressive component make up the bulk of the ARMA, while the ARIMA is an integrated moving average of autoregressive time series. When the time series is not stationary, the ARIMA model can be helpful. Moreover, the time series can't be considered stationary without the differencing[9].

When implementing the ARIMA model, we must take into account three values that are also parameters of the model. As a result, we can model it as (p, d, q).

P = lags in the autoregressive model.

D = differencing / integration order.

Q = moving average lags.

We need to specify two different kinds of orders in the SARIMAX model's parameter. For the first, we use a set of parameters (p, d, and q) that are analogous to those used in the ARIMAX model; for the second, we use a set of parameters (four numbers) to specify the impact of seasonality; this is known as a seasonal order.

Among the benefits of employing this algorithm are:

It is resilient against environmental factors. It's unique among models in this respect. Seasonal effects can be seen, for instance, in a time series where the temperature drops in the winter and rises in the summer. Even so, humidity's influence causes winter temperatures to rise, and rain brings the possibility of cooler conditions.

 https://doi.org/10.22362/ijcert/2022/v9/i11/v9i1103

If these factors don't show cyclical or seasonal behaviour, we won't be able to accurately predict their value. These data are outside the scope of other models' capabilities. The SARIMAX model is the only option if the data is not stationary[11].

4. Result and Analysis

For data collection we collected city data in days and hours from 2015–2020. Validated dataset against various environments 0074o see if it gives desired output.

Table 1. Attributes of the dataset

Attribute	Type	Description
City	String	Name Of the City
Date	Date	Date constituting pollutants
PM2.5	Numerical	Amount of PM2.5 present in air
PM10	Numerical	Amount of PM10 present in air
NO	Numerical	Amount of NO present in air
NO2	Numerical	Amount of NO2 present in air
NOx	Numerical	Amount of NOX present in air
NH3	Numerical	Amount of NH3 present in air
CO	Numerical	Amount of CO present in air
SO2	Numerical	Amount of SO2 present in air
O3	Numerical	Amount of O3 present in air
Benzene	Numerical	Amount of Benzene present in air
Toluene	Numerical	Amount of Toluene present in air
Xylene	Numerical	Amount of Xylene present in air

4.1 DATA PREPROCESSING

In the dataset there are two attributes that need to be preprocessed

- For the Xylene attribute, there are so many null values, if taken to consideration which in turn changes the precision value and has total count of 18109.
- For PM10 attribute, there are so many null values, if taken to consideration which in turn changes the precision value and has total count of 10328 as shown in table 2. So, thereby not taking them to consideration.

	No of missing values	% of missing values
Xylene	18109	61.320000
PM10	11140	37.720000
NH3	10328	34.970000
Toluene	8041	27.230000
Benzene	5623	19.040000
AQI	4681	15.850000
AQI_Bucket	4681	15.850000
PM2.5	4598	15.570000
NOx	4185	14.170000
O3	4022	13.620000
SO2	3854	13.050000
NO2	3585	12.140000
NO	3582	12.130000
CO	2059	6.970000

Table 2 Count of null values

4.2 DATA VISUALIZATION

The below are a Visualized graphs from the datasets.

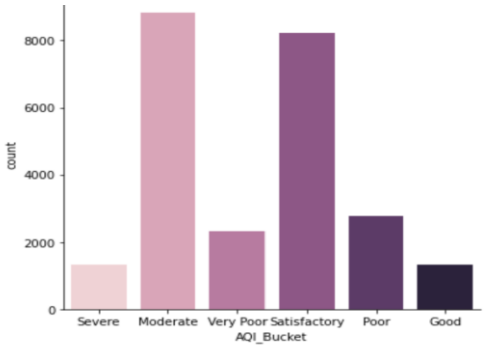

Figure 5: Bar plot displaying AQI Bucket against Count

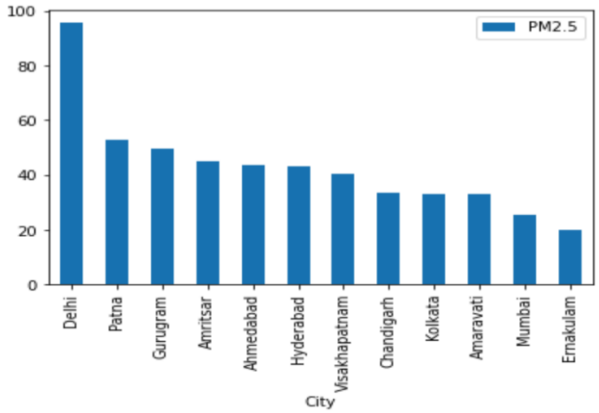

Figure 6: Bar plot displaying amount of PM2.5 present in every city.

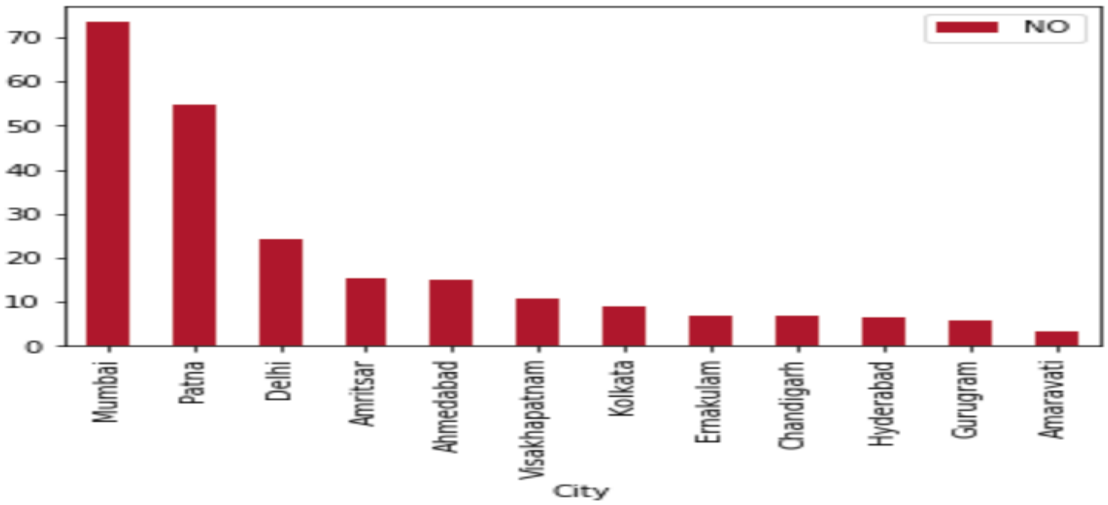

Figure 7: Bar plot displaying amount of Nitric oxide (NO) present in every city.

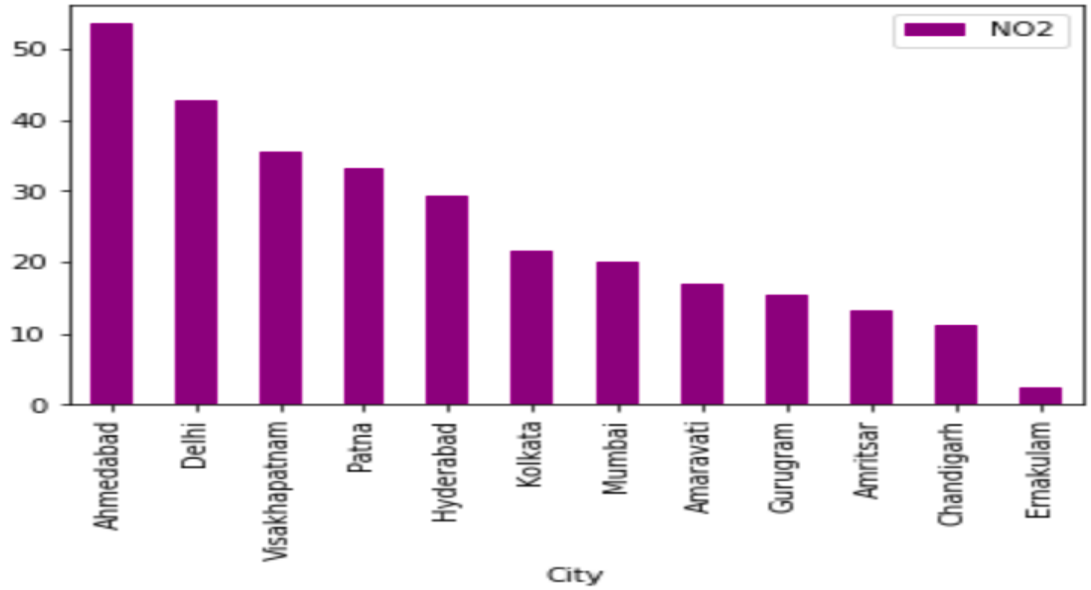

Figure 8. Bar plot displaying amount of Nitrogen Dioxide (NO2) present in every city

 https://doi.org/10.22362/ijcert/2022/v9/i11/v9i1103

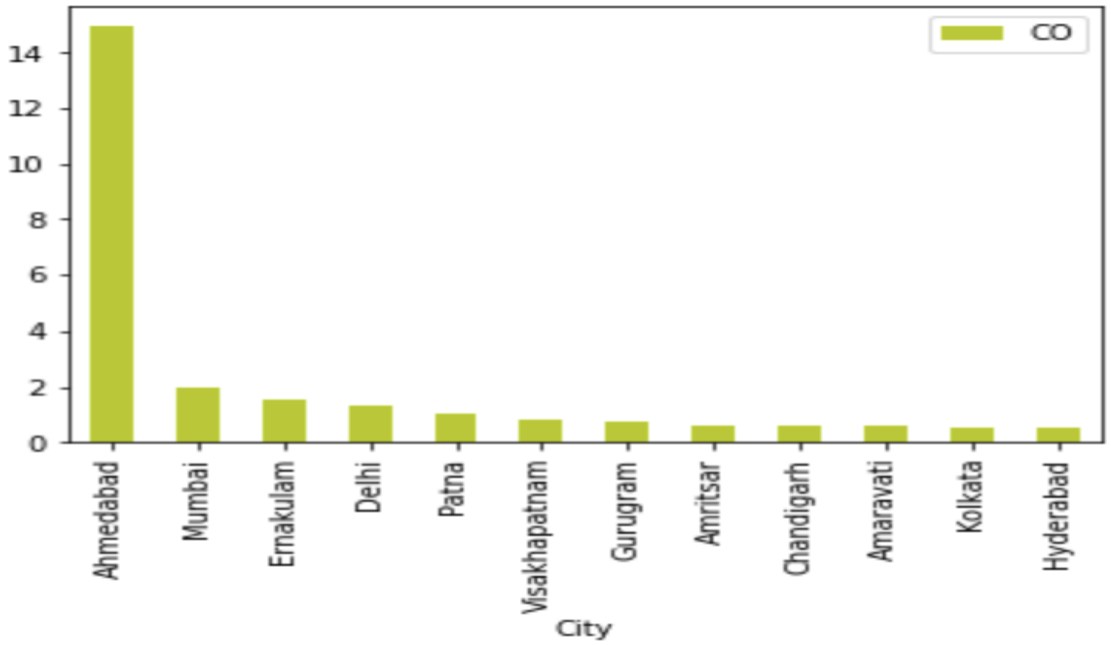

Figure 9: Bar plot displaying amount of Carbon Monoxide (CO) present in every city.

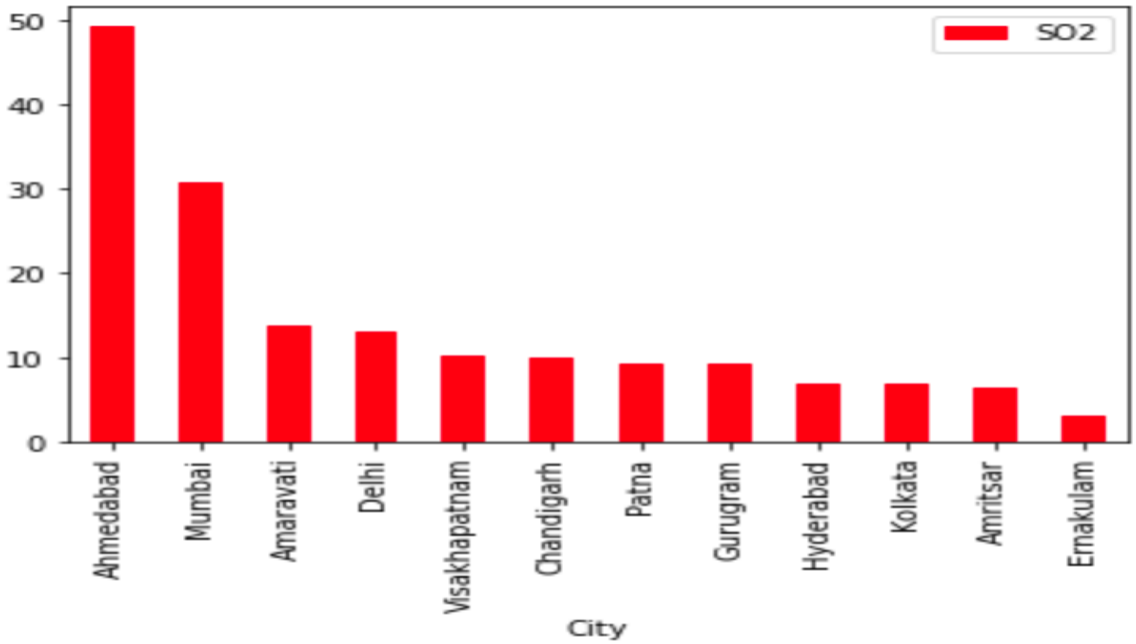

Figure 10: Bar plot displaying amount of Sulphur dioxide (SO2) present in every city.

 https://doi.org/10.22362/ijcert/2022/v9/i11/v9i1103

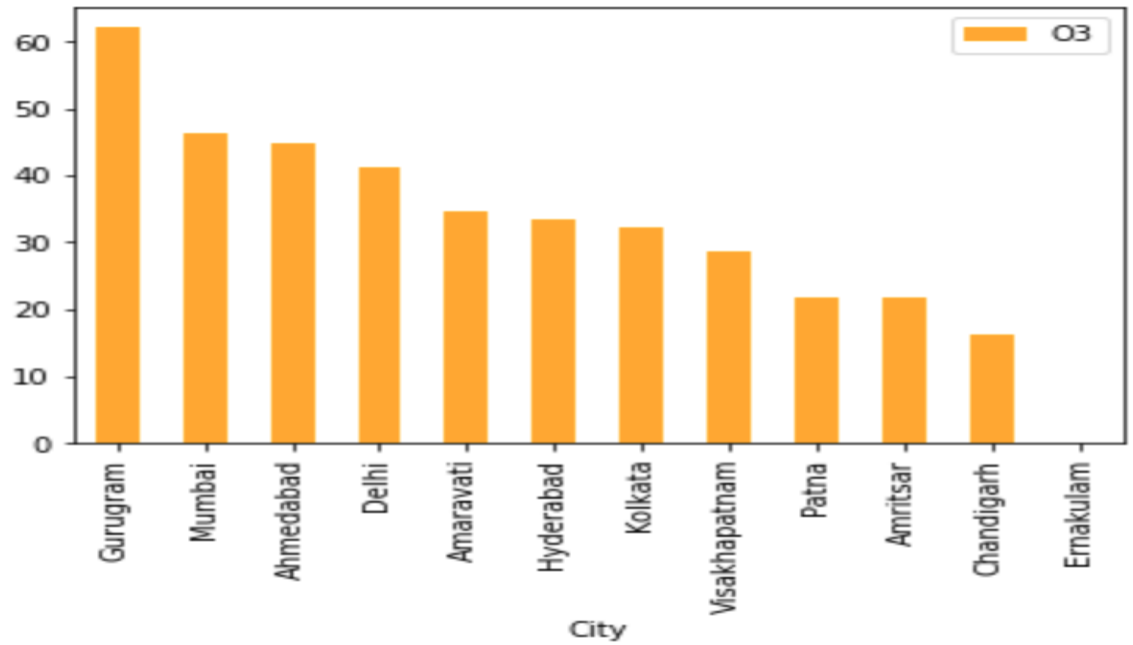

Figure11: Bar plot displaying amount of Ozone (O3) present in every city.

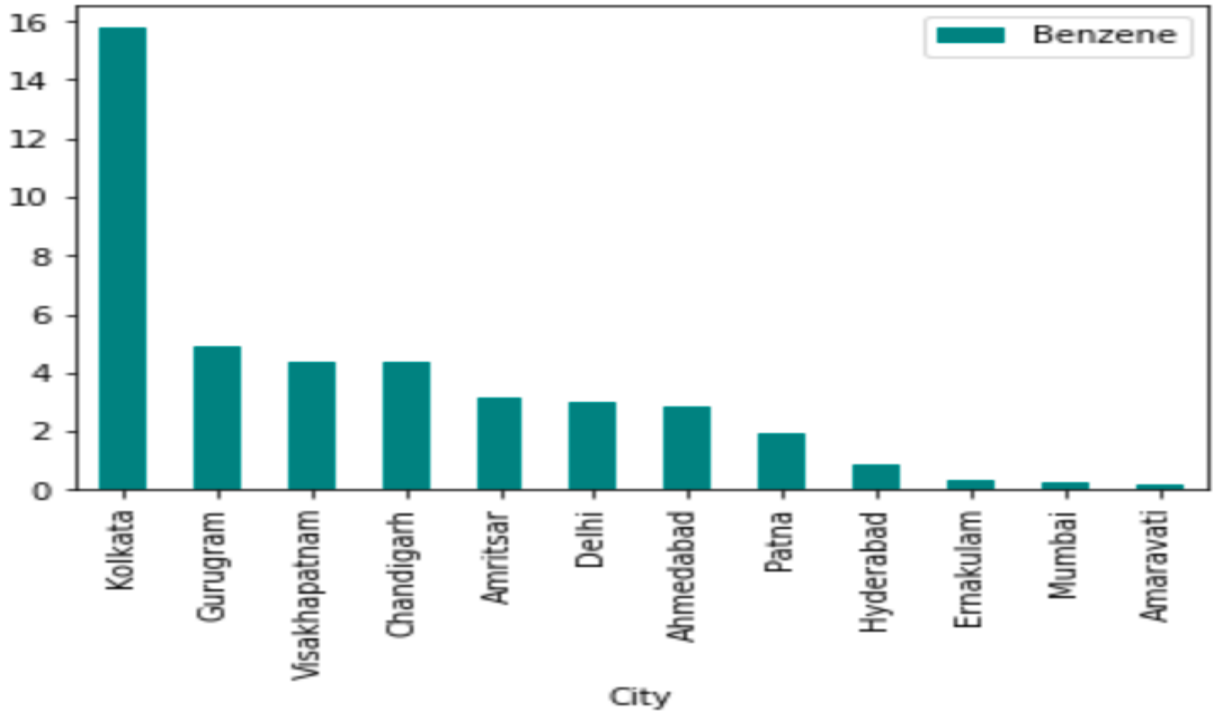

Figure 12: Bar plot displaying amount of Benzene present in every city.

These are eight different visualization in bar graphs each plotted against various pollutants.

4.3 FORECASTING

As we know that Forecasting is the use of a model to predict future values based on previously observed values. So, based on the data from 2015-2020 from various cities constituting plethora pollutants we first forecasted on past data to evaluate and check how precise the model works[12[.

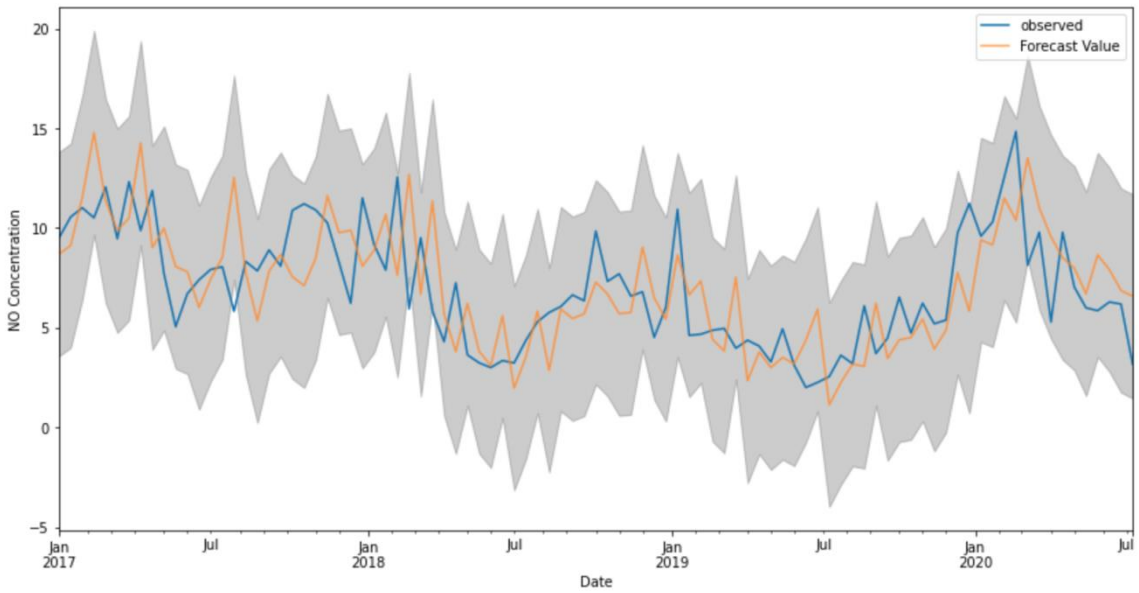

Figure 13: Evaluating data from 2017-2020 of pollutant Nitric oxide (NO)

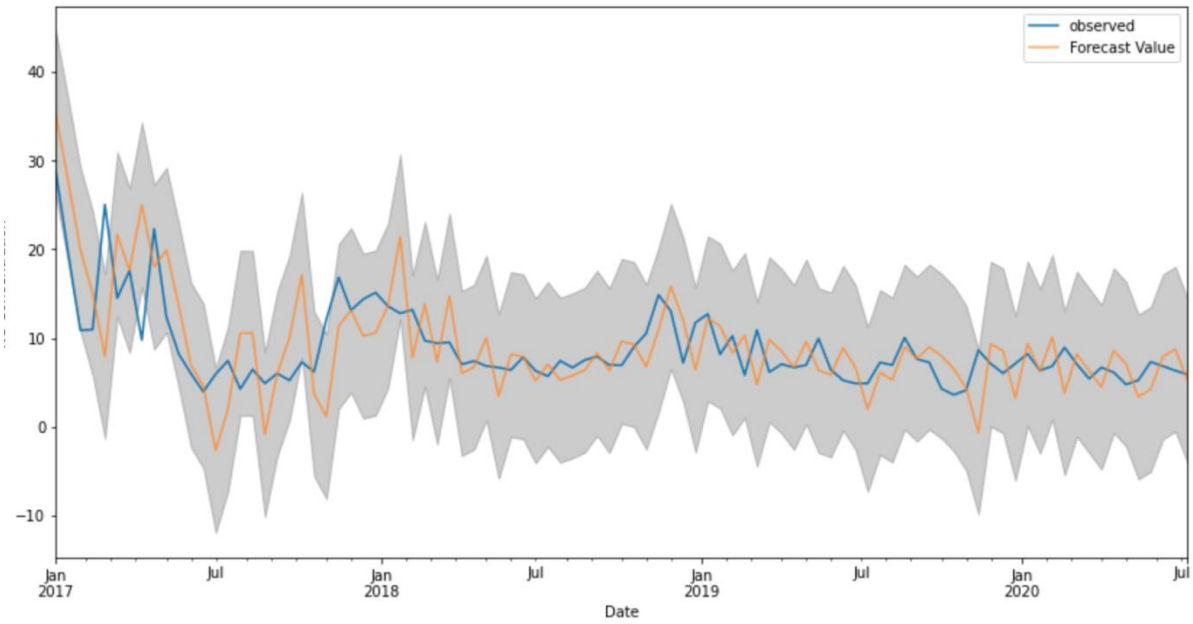

Figure 14: Evaluating data from 2017-2020 of pollutant Sulfur dioxide (SO2)

As the model gave precise results based on past data we have forecasted for next 3 years which we will discuss in coming chapter 7.

4.4 BINNING

Binning method is used to smoothing data or to handle noisy data. In this method, the data is first sorted and then the sorted values are distributed into a number of buckets or bins. Smoothing by bin median : In this method each bin value is replaced by its bin median value. Binning: Binning into 6 buckets based on AQI[13] values output:

```
Moderate        71987
Satisfactory    63724
Good            23666
Poor             2800
Very Poor        1029
Severe            397
Name: AQI_bucket_calculated, dtype: int64
```

Figure 15: Represents AQI severity based on Each Value

5. Conclusion

If the historical data is any indication, nitric oxide levels are expected to decrease during the next few years. However, Figure 7.2 shows that SO2 is expected to increase relative to historical averages. Pollution control in Telangana estimates that there are 1.38 crore automobiles in the state, of different types. Smoke from vehicles is a growing source of air pollution and a health threat as the number of cars on the road continues to rise. Collecting information from coal-fired power plants and rural India would be our first priority. Find the primary pollutant emitted by coal-fired power plants and perform binning on the waste. Based on the historical data, projections are made for the following years. Since India, the third-largest producer of greenhouse gases in the world, has committed to reaching net-zero carbon emissions by 2070, we will provide recommendations based on the impact.

References

[1] Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019.Open AccessPublished:December 21, 2020DOI:https://doi.org/10.1016/S2542-5196(20)30298-9

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465283

[3] A Study and Analysis of Air Quality Index and Related Health Impact on Public Health

[4] Study of the Effects of Air Pollutants on Human Health Based on Baidu Indices of Disease Symptoms and Air Quality Monitoring Data in Beijing, China.

[5] A. Kumar and P. Goyal, "Forecasting of air quality in delhi using principal component regression technique," Atmospheric Pollution Research, vol. 2, no. 4, pp. 436–444, 2011.

[6] C. Zhang, J. Yan, Y. Li, F. Sun, J. Yan, D. Zhang, X. Rui, and R. Bie, "Early air pollution forecasting as a service: An ensemble learning approach," in Proc. IEEE Int. Conf. on Web Services (ICWS), 2017, pp. 636–643.

[7] Z. Wang and Z. Long, "Pm2. 5 prediction based on neural network," in Proc. IEEE 11th Int. Conf. on Intelligent Computation Technology and Automation (ICICTA), 2018, pp. 44–47.

[8] K. B. Shaban, A. Kadri, and E. Rezk, "Urban air pollution monitoring system with forecasting models," IEEE Sensors Journal, vol. 16, no. 8, pp. 2598–2606, 2016.

[9] B. Yeganeh, M. S. P. Motlagh, Y. Rashidi, and H. Kamalan, "Prediction of co concentrations based on a hybrid partial least square and support vector machine model," Atmospheric Environment, vol. 55, pp. 357–365, 2012

[10] Kurt, A. and A.B. Oktay, 2010. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. Appli., 37: 7986-7992. DOI: 10.1016/j.eswa.2010.05.093

[11] Yedukondalu, Gangolu & K., Samunnisa & Bhavsingh, M. & Raghuram, I & Lavanya, Addepalli. (2022). MOCF: A Multi-Objective Clustering Framework using an Improved Particle Swarm Optimization Algorithm. International Journal on Recent and Innovation Trends in Computing and Communication. 10. 143-154. 10.17762/ijritcc.v10i10.5743.

[12] Maloth, Bhav Singh & Anusha, R. & Reddy, R. & Devi, S.Chaya. (2013). Augmentation of Information Security by Cryptography in Cloud Computing. www.ijcst.com. 4.

[13] Samunnisa, K. & Kumar, G. & Madhavi, K.. (2022). Intrusion detection system in distributed cloud computing: Hybrid clustering and classification methods. Measurement: Sensors. 25. 100612. 10.1016/j.measen.2022.100612.

