Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-3 Issue-4, June 2023

Is the Ecosystem of Kolkata Sustainable?: Machine Learning Based Study on Air Quality Index

Biswajit Biswas, Sayantan Ghosh

Abstract: ***Timely and accurate forecasting of Air Quality Index (AQI) helps the Industries to select suitable control of air pollution measures. It helps people to reduce exposure in pollution. In this present age Air quality Index is one of the burning issues in India. The air contaminations are harmful for our biological system and also for the climate. To keep up the best air quality cross the country different types of air toxins are estimated through the air quality measuring standards. The aim of this research work is modelling air quality of a location with respect to time with the help of Machine Learning (ML). The proposed and developed model was emphasizes particularly in Kolkata, capital of the state West Bengal in India and the findings have direct implications to build & maintain a sustainable ecosystem over there.***

Keywords: Air Quality Index, Machine Learning (ML), Python, Suitable control.

I. INTRODUCTION

Air pollution caused by the presence of compounds in the atmosphere that are hazardous to human and other living animals as well as to the environment also. Gases, particles, biological molecules and various types of air contaminants are present in the mixture of air pollutants. Air pollution has become a serious matter in many of the urban cities in India. Every human being must know about the air quality that they are breathing, so CPCB had developed the Air Quality Index (AQI) for every city in India. The AQI gives an idea on the quality of air for that particular area whether that area is polluted or not. AQI is a numeric value by which government pollution board measure the air pollutants level present in the atmosphere. If the AQI value is increased then percentage of pollutants is high and that can affect adversely human health. According to Central Pollution Board there are twelve parameters present in the air pollutant and author have taken the most important pollutants which are very harmful to atmosphere like $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, OZONE.
The selection of these pollutants varies on availability of AQI, data frequency and measurement methods. In respect to CPCB the AQI gives an idea about the air quality to what extent the particular area is polluted which gives an idea that AQI provides the actual value of Air Quality in our eco system which is in touch with our different health issues.

Manuscript received on 28 April 2023 | Revised Manuscript received on 01 June 2023 | Manuscript Accepted on 15 June 2023 | Manuscript published on 30 December 2023.
** Correspondence Author (s)*
Dr. Biswajit Biswas*, Department of Business Administration, University of Kalyani, West Bengal, India. E-mail: biswajit.biswas0012@gmail.com, ORCID ID: 0000-0002-5302-7675
Sayantan Ghosh, Performance-io LLP, Kolkata (West Bengal), India. E-mail: sayantang28@gmail.com

In this work authors focused on the AQI prediction models and took six months data of Kolkata region where Ballygunge is the focused area of testing among other centers. This work is being performed in an open source platform 'Python' using Jupyter Anaconda and also authors imported packages for predictions and data processing those are Numpy, Pandas, Matplotlib.pylot, Seaborn, Sklearn, Warnings.

II. OBJECTIVE OF THE STUDY

a. To design a prediction model for forecasting the air quality.
b. To gather a brief knowledge on Air Quality Index and know about the bad impacts of air pollutants that is affecting adversely on human health.

III. LITERATURE REVIEW

In this paper he had taken hourly pollutant level concentration data of Canada. He had used accuracy, efficiency and up ability as key indicators in his research work. He had used machine learning methods using Extreme Learning Machine (ELM) which is an updated algorithm and different from other models used for forecasting. (Meng Dun et al. 2020) [1])

They had research on complete data analysis of pollutants and their model can forecast AQI of any region with more than 90% accuracy. In their study the author had taken China AQI at the time of checking and investigating the data. They took each air pollutants concentration and their percentage of impact in AIR using Gradient Boost Algorithm (Akshaya A.C et al. 2019[2]) Now a day Air Pollution is a serious matter for the human life and it leads to premature death. Researchers try to predict the AQI to alert people in advance for their living surroundings. AQI measure depends on NO_2, CO, SO_2, PM_{10}, $PM_{2.5}$, NH_3, 0_3 and Pb, among them $PM_{2.5}$ is too much affecting in human life. In this study authors gave major focus on $PM_{2.5}$. (J.K.Sethi and M.Mittal, 2019[3][12][13][14][15][16]) The main agenda of this research paper is to discover the AQI model and to find out the impacts of polluted air in life of every human being. In this paper they had first applied calculation of AQI by taking the concentration of different air pollutants and getting a single numerical value which is known as AQI value. Then the authors of this paper had done aggregated Index Calculation which helps in finding the air quality condition and its impact on our environment. According to this study, I found that they generally worked on social experiment of AQI by using ANN, Linear & Logistic Regression and how air pollution is affecting our lives (Radhika M Patilet al. 2020[4]).

In this study authors took data over eleven years from three regions in Taiwan.

They applied two machine learning method for AQI prediction which is new and give better results from other machine learning algorithms. They created a model which illustrates analysis and prediction of Air Quality System (Yona Maimuryet al., 2020[5]). In this work, authors created a hybrid model using multivariable regression and support vector regression to predict the pollutants present in air. Authors claimed that their hybrid model gives better accuracy percentage than the other single model present in the market for prediction of AQI. In this study authors created a model using Data tree, SVM, KNN, RF and Logistic Regression which gives a daily observation of the air pollutant and their accuracy is not less than 92 percent as per Ministry of Environment. In this paper they had find out the most contaminated sites and the pollutant concentration present there and researched on it how to make the air pollutant free to clean ecosystem (Khalid Naharet al., 2020[6]). In this civilization age population is increasing day by day. So it needs to forecast the air quality for early warning about the air pollutants those are harmful for the human life also for the ecosystem. In this work authors done a statistical analysis and predict the pollutants present in the air. This study done based on the historical data in Malaysia, (W.Y. Hong et al., 2021[7]). Air quality measurement is a challenging area for the present researchers because of its harmful impact on the ecosystem. The AQI is increasing due to $PM_{2.5}$ that affect human lung, kidney, brain liver and it lead to cancer. Authors developed a frame work for observing the air quality using ML (Machine Learning) algorithm. (R. Sharma, 2021[8]). In this study authors had used Gaussian Naive Bayes model which has the highest accuracy in terms of forecasting than Support Vector Machine model. They had also created model with XG Boost which has the highest linearity than actual and predicted data, (K. Kumar and B.P. Pande, 2022[9]). In this research work authors had taken the data of Delhi AQI. They used Support Vector Regression (SVR) model for forecasting the pollutants levels present in the AQI of Delhi. Along with SVR they also used RBF (Radical Basis Function) which gives better results in forecasting and analyzing the data. With the help of this analysis there model predicts various pollutant level with an accuracy of 93.4 percent, (S. Bhattacharya and SK Shahnwaz [10]).

IV. RESEARCH GAP

From the literature review, authors found that most of the author developed model for forecasting of present data available in the AQI website. There is no suitable model with web-based terminal where prediction of AQI can be done with the help of Predictive Model using Python libraries.

V. RESEARCH METHODOLOGY

In this work authors performed the statistical analysis by taking the parameters as Ozone, NO_2, SO_2, $PM_{2.5}$, PM_{10} which gives an idea about the awareness of pollutant that how much it is affecting our ecosystem. It is being performed in Jupyter Notebook with the help of machine learning techniques. The details of this analysis are described in below.

A. Box-Plot Analysis

It is used in graphical representations of numerical information of data. It contains lines which divides data set in form of three quartiles which represents minimum, maximum, median, first quartile and third quartile. In this work with the help of the analysis it created a graph with x-axis and y-axis that visualizes the value distribution in respect with each variable (air pollutants) taken on this data. The box is called interquartile range (IQR) the middle line of the box is defined as median and upper point is denoted as lower quartile Q1 and lower point is said as upper quartile Q3 (D. Meng et al., 2020[11]).

B. Heatmap

It is used to show two-dimensional graphical representations of data in which individual values are represented in form of colors contain in a matrix. It works with the help of Seaborn package that is been imported in this work. Here, the pollutants with higher values are represented in darker shades and lower values in lighter shades.

C. Pair Plots

Pair plots help in creating axes in which each variable presents in the data in a way that x-axis and y-axis are directly in proportional to each other as row and column that creates a relationship in a regressor format among the dataset.

D. Scatter Plot

It helps in this work for observing correlation in between variables and utilizes dots to identify the relationship among them. It plots data points in horizontal axis and vertical axis to display how each variable is connected to another.

E. Matplotlib Histogram

It is used in visualizing the frequency distribution by separating the array (numeric) to mini same sized bins. Then the variable forms a continuous distribution program which is very helpful to compare by various classifications. It is the main skill which is used in data science for building a frequency table that are generally taken from a complete dataset to easy to learn the various elements occurrence and it is the main purpose of Matplotlib Histogram package used by axes subplot.

F. Linear Regression

It is one of the best used regression techniques performed in machine learning. It performs statistical analysis which creates model that builds a relationship with an independent variable and dependent variable. It helps in predicting a response whether the two variables are linearly related or not. The author tried to find a prediction by taking response value from dataset(y) and taking independent variable(x) which will give a graphical analysis that what can be the condition of air quality in future.

VI. DATA ANALYSIS AND INTERPRETATION

The python packages are used in this model to import data analysis and prediction purpose.

In Fig .1 have taken numPy package for mathematical operations, pandas for data analysis, matplotlib.pylot package for plotting histogram and scattering plot, seaborn package is taken for visualization of data.

```
In [4]:
import numpy as np # linear algebra
import pandas as pd # data processing

import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
plt.rcParams['figure.figsize'] = (10, 7)
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Warnings
import warnings
warnings.filterwarnings('ignore')
import pandas.util.testing as tm
```

Figure. 1: Package Import

	S.No	From Date	To Date	PM2.5 (ug/m3)	PM10 (ug/m3)	NO2 (ug/m3)	SO2 (ug/m3)	Ozone (ug/m3)	CO (mg/m3)
0	1	01-Jan-2022 - 00:00	02-Jan-2022 - 00:00	109.44	180.29	32.75	29.09	39.41	0.92
1	2	02-Jan-2022 - 00:00	03-Jan-2022 - 00:00	106.82	163.64	30.43	23.51	35.74	0.85
2	3	03-Jan-2022 - 00:00	04-Jan-2022 - 00:00	115.89	180.47	35.23	20.00	34.92	1.20
3	4	04-Jan-2022 - 00:00	05-Jan-2022 - 00:00	149.07	221.46	36.98	15.81	40.85	1.18
4	5	05-Jan-2022 - 00:00	06-Jan-2022 - 00:00	162.12	242.20	40.35	21.18	44.60	1.60

Figure. 2: Data Import

	S.No	PM2.5 (ug/m3)	PM10 (ug/m3)	NO2 (ug/m3)	SO2 (ug/m3)	Ozone (ug/m3)	CO (mg/m3)
count	150.000000	150.000000	150.000000	119.000000	150.000000	130.000000	150.000000
mean	75.500000	57.105333	114.496200	35.220756	13.167200	48.874846	0.720267
std	43.445368	37.307898	60.127586	28.099272	8.203153	13.925851	0.347132
min	1.000000	5.950000	33.270000	6.610000	1.030000	20.470000	0.210000
25%	38.250000	25.195000	63.992500	8.870000	6.862500	39.987500	0.460000
50%	75.500000	51.240000	96.795000	29.390000	12.120000	47.285000	0.675000
75%	112.750000	84.452500	168.207500	51.035000	18.565000	56.032500	0.877500
max	150.000000	162.120000	257.270000	170.070000	46.050000	99.440000	2.020000

Figure. 3: Data Description

In Fig.2 the dataset is imported and viewing the shape of dataset whether the data is imported properly or not for analysis. In Fig.3 the data is described and displayed a statistical summary of the data frame. If the dataset is containing a numerical value then data. Describe () command is used for graphical representation which measures description of the data in form of descriptive statistics. The Box-plot analysis is done in Fig.4 to show how the data is well distributed in the dataset as it is a type of chart which explains visual distribution of numerical data displaying in form of quartile data.

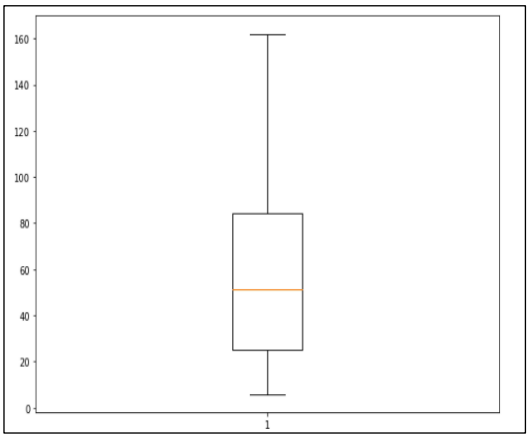

Figure. 4: Box-Plot Analysis

With the help of panda's visual analysis scattered plot of Ozone is displayed here in Fig.5 by collecting the pairs of data in which a relationship is identified. Then the graph is drawn with independent variables on horizontal axis and dependent variables on vertical axis.

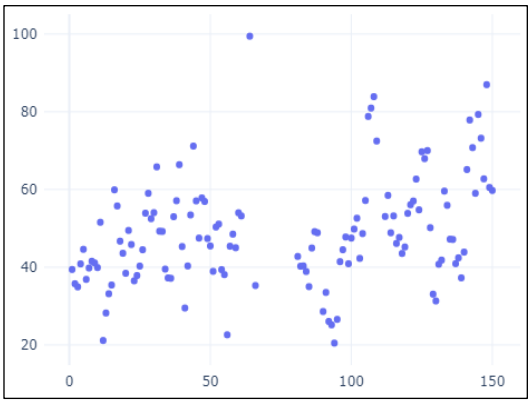

Figure. 5: Scattering Plot

The histogram of PM 10 is shown in Fig.6 using visual analysis which shows an accurate display of distribution of numerical data and range of values are divided into series of intervals.

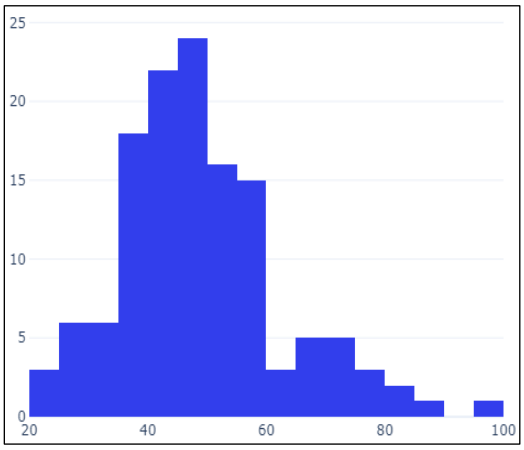

Figure. 6: Histogram

In this Fig.7 the scattered plot is analysed by taking two axes where x-axis is NO2 and y-axis is SO2 which perform a comparison between the variables and identifying the common variables and distributed in form of scattering format.

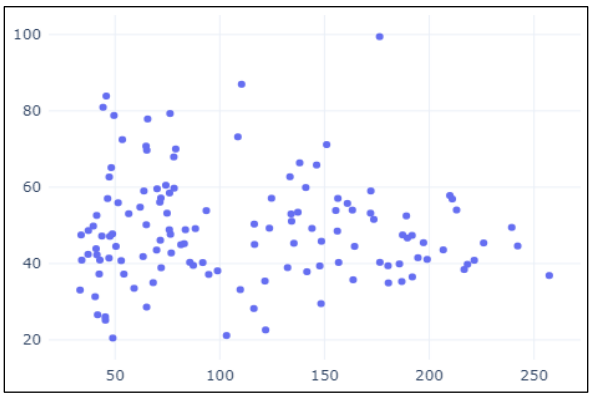

Figure. 7: Scattered Plot

Here the heatmap analysis is performed in Fig.8 where two-dimensional graphical representation is shown in form of matrix that are represented by colours and it performs two-dimensional plot where values are mapped on indices and columns to the chart.

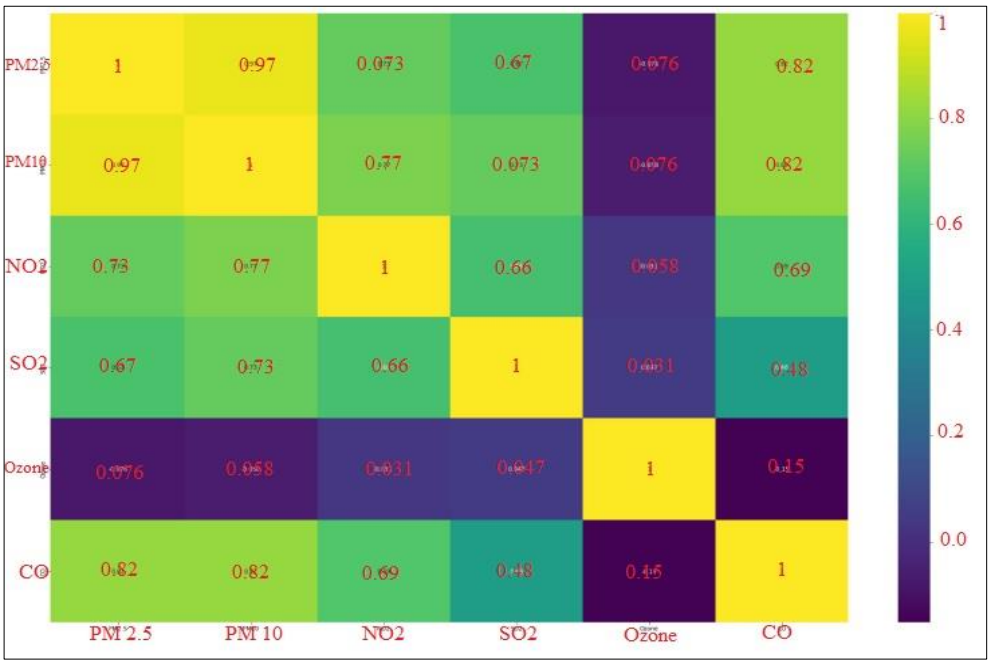

Figure. 8: Heatmap Analysis

With the help of Seaborn package pair plot function is performed here in Fig.9 which helps to understand the relationship among each variable. Scattering plot of PM2.5 is shown here in Fig.10

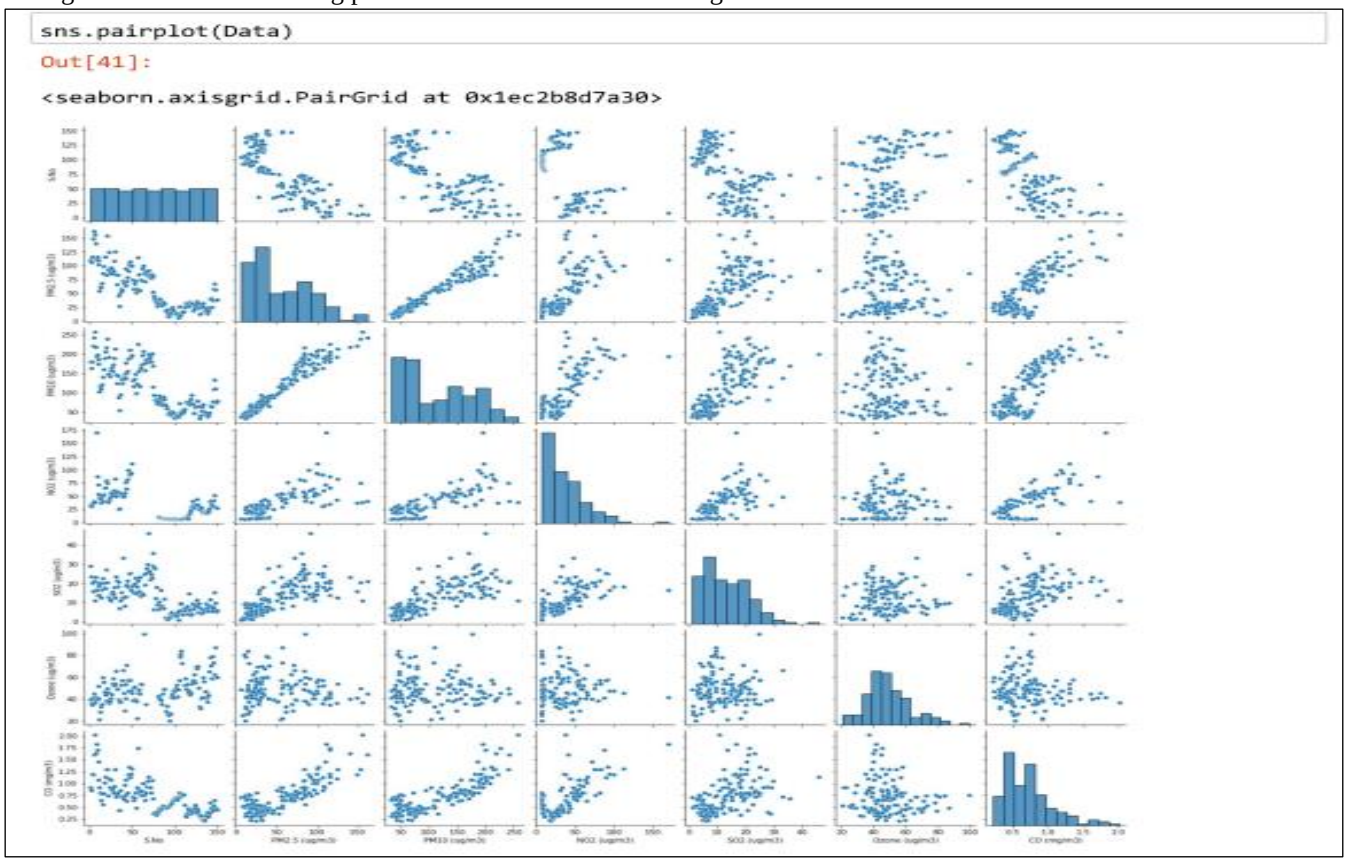

Figure. 9: Pair Plots Analysis

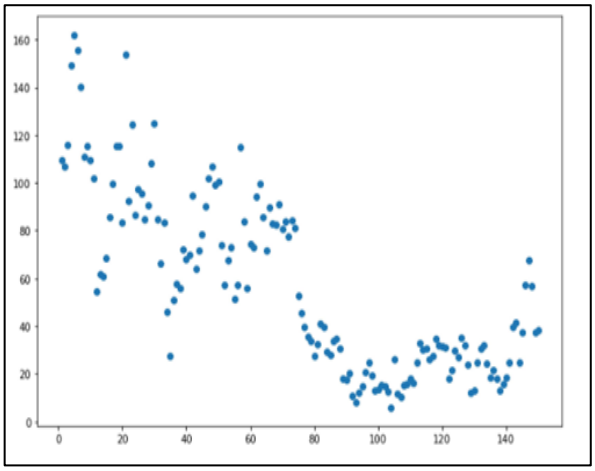

Figure. 10: Scattering Plot of PM 2.5

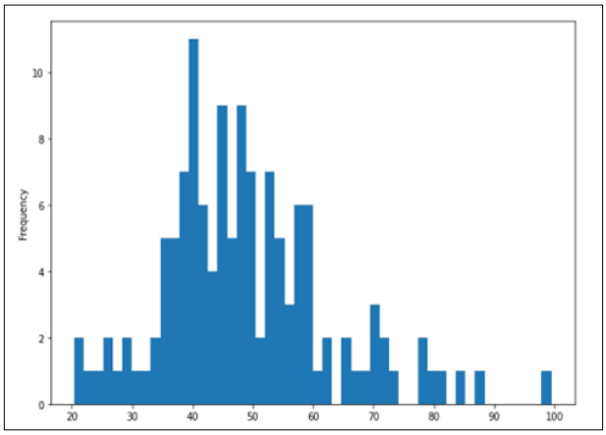

Figure. 11: Frequency Check of Ozone

In this histogram the frequency level of Ozone is checked and shown in Fig.11.After the train and test of data now it is the time for predicting the pollutants, author have taken the data of five months in which three months data have been taken as an actual data and two months data is used for prediction purpose. In Fig .12 we have taken the data PM10 in respect to time where the blue line is identified as an actual data and orange line as the predicted data. As the two months data is already in our hand from before so manually while checking the predicted data with the original data set it seems that our prediction is 90% accurate and from here author can say that our prediction model is quite satisfactory and performing well.

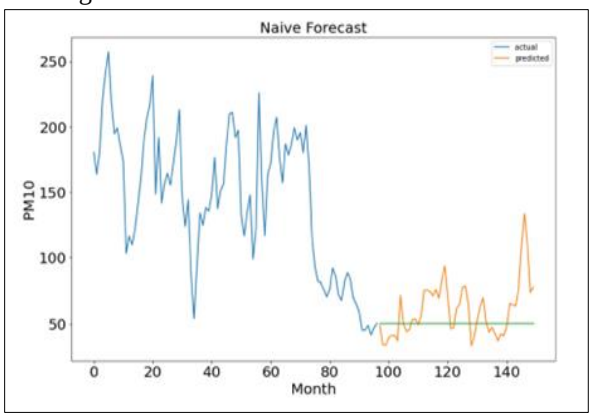

Figure. 12: Actual vs. Predicted Data of PM10

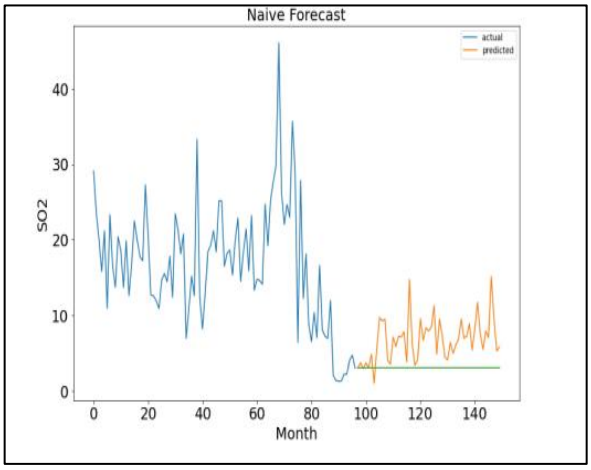

Figure. 13: Actual vs. Predicted Data of SO2

In Fig. 13 the prediction is done on SO2 as the same process which is done on previous prediction model and it gave us the 90% accuracy according to our dataset. From here we can say that our model is performing well.

VII. MANAGERIAL IMPLICATIONS

1. In this work, author is building a hybrid model using machine learning algorithms which will help in predicting the AQI in contaminated areas/polluted cities.
2. From this model a word of pollutant control will work if the prediction is done correctly using proper algorithms.
3. Basically the model will help in predicting the AQI and making an environment pollutant control and creating a clean ecosystem.

VIII. LIMITATIONS

1. As the data of AQI is taken from government site author worked with the static data only but if that is done with real time data using cloud computing it could give better result.
2. As the analysis of AQI is done with Kolkata which intake only less data, but for future work while working with large amount of data scope of using Genetic Algorithm is a better method.
3. The model is created with low amount of data so complexity is very high which will create impact on prediction analysis.
4. Manufacturing industries can know about the harmful partial those are exhausted from their industries.

IX. FUTURE SCOPE

1. This model can be extended in future by using Deep Learning Techniques for better accuracy.
2. A web based terminal can be created for prediction of AQI using cloud computing where real time data will be used for forecasting.
3. A model needs to be created which will not only work on predicting the air pollutants but also on meteorological parameters also for analyzing the concentration level.
4. In this model we are using Machine Learning Algorithms but if the prediction is done with Artificial Neural Networks the prediction results will be more prominent better than this model.

X. CONCLUSION

The purpose of this work is to understand the Air Quality Index (AQI) and know about the harmful particulars present in the air. The work of predicting the pollutant levels is quite tough. However, the tasks of predicting air in this work authors used Linear Regression for predicting the levels of pollutants like $PM_{2.5}$, PM_{10}, NO_2, SO_2, Ozone and the data are collected from the CPCB website. With the help of this model it may know the surrounding environment. The Model maybe enhanced with other Machine Learning Algorithm.

DECLARATION STATEMENT

In this research work authors used publicly available data across Ecommerce websites. This work does not contain any studies with human participants or animals performed by any of the authors. Further, all authors are attached in the Department of Business Administration, University of Kalyani and consequently they used the infrastructure of their university to carry on this research activity. Consequently, there is no conflict of interest involved in this case. There is no internal and external funding agency to complete this research work.

Funding	No, did not receive fund from any resources.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	In this research work authors used publicly available data across Ecommerce websites.
Authors Contributions	All authors having equal participation in the article.

REFERENCE

1. Meng Dun (et al. 2020) 'Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine' Volume 2020 |ArticleID 8914501 | https://doi.org/10.1155/2020/8914501 https://doi.org/10.1155/2020/8914501
2. Akshaya A.C(etal.2019) 'Indian Air Quality Prediction and Analysis using Machine Learning Volume 14, Number 11, 2019 (Special Issue) © Research India Publications. http://www.ripublication.com
3. J.K. Sethi and M.Mittal(2019), 'A new feature selection method based on machine learning technique for air quality dataset', Journal of Statistics & Management SystemsVol.22(2019), No.4, pp.697-705, DOI:10.1080/09720510.2019.1609726, @ Taylor & Francis. https://doi.org/10.1080/09720510.2019.1609726
4. Radhika M Patil (et al. 2020) 'Prediction an air quality index data using machine learning and deep learning' http://norma.ncirl.ie/5208/1/ruchitadattatraypatil.pdf
5. Naama Lang-Yona, Fatma Öztürk, Daniella Gat, Merve Aktürk, Emre Dikmen, Pavlos Zarmpas, Maria Tsagkaraki, Nikolaos Mihalopoulos, Aşkın Birgül, Perihan Binnur Kurt-Karakuş, Yinon Rudich, Links between airborne microbiome, meteorology, and chemical composition in northwestern Turkey, Science of The Total Environment,Volume 725,2020,ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.138227
6. Khalid Nahar (et al.2020) ' Air Quality Index Using Machine.
7. W.Y.Hong, D.Koh and A.A.A.Mohtar (2021)," Statistical Analysis and Predictive Modeling of Air Pollutants Using Advanced Machine Learning Approaches". Asia-Pacific Conference on Computer Science and Data Engineering (CSDE 2020) @IEEE, https://doi.org/10.1109/CSDE50874.2020.9411636

8. R.Sharma, G.Shilimkar and S.Pisal (2021),"Air Quality Prediction by Machine Learning", International Journal of Scientific Research in Science and Technology, ISSN:2395-6011, https://doi.org/10.32628/IJSRST218396
9. Kumar K, Pande BP. Air pollution prediction with machine learning: a case study of Indian cities. Int J Environ Sci Technol (Tehran). 2023;20(5):5333-5348. doi: 10.1007/s13762-022-04241-5. Epub 2022 May 15. PMID: 35603096; PMCID: PMC9107909. https://doi.org/10.1007/s13762-022-04241-5
10. S. Bhattacharya and SK Shahnwaz 'Using Machine Learning to Predict Air Quality Index in New Delhi' https://arxiv.org/ftp/arxiv/papers/2112/2112.05753.pdf
11. D.Meng-Chuen Chen *et al* 2020 *Environ. Res. Lett.* 15 074021DOI 10.1088/1748-9326/ab8659
12. Nikam, S. S., & Dalvi, Prof. R. (2020). Fake News Detection on Social Media using Machine Learning Techniques. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 7, pp. 940–943). https://doi.org/10.35940/ijitee.g5428.059720
13. Radhamani, V., & Dalin, G. (2019). Significance of Artificial Intelligence and Machine Learning Techniques in Smart Cloud Computing: A Review. In International Journal of Soft Computing and Engineering (Vol. 9, Issue 3, pp. 1–7). https://doi.org/10.35940/ijsce.c3265.099319
14. Dogra, A., & Dr. Taqdir. (2019). Detecting Intrusion with High Accuracy: using Hybrid K-Multi Layer Perceptron. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 3, pp. 4994–4999). https://doi.org/10.35940/ijrte.c5645.098319
15. Chandrababu, M., & Moorthy, Dr. S. K. K. (2022). Proficient Machine Learning Techniques for a Secured Cloud Environment. In International Journal of Engineering and Advanced Technology (Vol. 11, Issue 6, pp. 74–81). https://doi.org/10.35940/ijeat.f3730.0811622
16. Sharma, P., & Site, S. (2022). A Comprehensive Study on Different Machine Learning Techniques to Predict Heart Disease. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 2, Issue 3, pp. 1–7). https://doi.org/10.54105/ijainn.c1046.042322

AUTHORS PROFILE

Dr. Biswajit Biswas is an Assistant Professor in the Department of Business Administration, University of Kalyani. He has completed his B.Tech and MBA in Information Technology from the University of Kalyani, India. He has got first class throughout his academic. He has qualified the UGC-NET and awarded Ph.D from the University of Kalyani, India. Broad area of his Ph.D. is the Business Intelligence Model in Digital Marketing Using Soft Computing Approaches. He has more than seven years research and teaching experience in Post Graduate and Under Graduate level in the domain of Business Administration. His expertise is about Business Intelligence, Web-data mining, Soft computing, Principle of Management, Strategic Management and Digital Marketing. He has published many research articles in International journal and conferences. He also visited many prestigious academic Institutes like IIMA, IIMB, IIMC, IISC, AMITY University, KIIT, DUYTUN University Vietnam and many more in India and aboard. He has attached several teaching institutes including undergraduate & postgraduate levels. He is a reviewer of many prestigious International Journals.

Mr. Sayantan Ghosh is a Technical SEO Associate at Performance-io India, leveraging his expertise in the intersection of technology and marketing. With a MBA in Information Technology from Kalyani University and a Bachelor's in Computer Science and Engineering from Chandigarh University, Mr. Ghosh has cultivated a deep understanding of both business administration and technical facets in the digital realm. Beyond his IT proficiency, Mr. Ghosh holds a profound fascination for the Stock Market, fostering an avid interest in analyzing market trends and predicting future trajectories.