Effect of Covid-19 on Air Quality Index

Manish Sharma
Computer Science Engineering
Bennett University
Greater Noida, India
e19cse053@bennett.edu.in

Karthikeyan Rathore
Computer Science Engineering
Bennett University
Greater Noida, India
e19cse119@bennett.edu.in

Tanuj Sharma
Computer Science Engineering
Bennett University
Greater Noida, India
e19cse147@bennett.edu.in

***Abstract*—At the start of 2020, we began to see a rise in concern regarding the COVID 19. On 11th March 2020, the COVID 19 is declared a pandemic by WHO. The presented study aims to see visualize the changes in the Air Quality Index(AQI) that were caused by the nationwide lockdown because of the COVID 19. The study will compare the state of air quality index in India before and after the global wide lockdown. The research data used was collected and organized by the authorized body of the Indian government (Central Pollution Control Board). The information outline begins from 2015 to 2020 and consists of 16 columns which include 1 target column with the Air Quality Index and 15 features like PM2.5, Ozone, Benzene, etc. For data preprocessing we will be using Simpleimputer, which will be providing various basic techniques for substituting the missing values. We compared different models to see the best predicting one. We hyper tuned the models to find the best parameters for it. And using various evaluation matrices, we conclude that the RandomForestRegressor and Sarima were suited for predicting Air Quality Index in India.**

I. Introduction

History has seen many epidemics and pandemics arise and then die out, as we evolve further, so do the viruses. The recent novel coronavirus (COVID-19) outbreak which spread in late 2019 from Wuhan, China had forced us to go into lockdown [1]. As this emerged, the complete lockdown of industries and cities took place and researchers and scientists from all over the world started evaluating the short and long-term effects of the global pandemic. Despite all the negative effects including new social norms and the use of face masks at all times, this obligatory lockdown during the pandemic created a positive impact on the environment [2]. With the discontinuation of all significant activities like restricted vehicle movements and the shutdown of industries, the number of air pollutants decreased drastically worldwide which resulted in an improvement of the Air quality index (AQI), which had previously been thought impossible despite multiple efforts.

Particulate Matter often abbreviated as PM refers to anything in the air that is not a gas. These are very microscopic particles suspended in the air we breathe and comprise of nitrates, sulfate, sodium chloride, nitrates, black carbon, mineral dust and water. Although particles matter with diameter 10 micrometers can easily penetrate deep inside of human lungs, particles matter with diameter 2.5 micrometers or less (PM 2.5) can possess a serious threat to health. Not only they can effortlessly infiltrate the lung barrier but also enter the human blood stream. All epidemiological and related evidence suggests exposure to these fine particles as the major cause of cardiopulmonary disease [3], lung cancer [4] and infant mortality [5]. Air pollution alone was responsible for 6% of total mortality or more than 40,000 fatalities each year. Out of the total fatalities due to air pollution, motorized vehicles contribute about half (50%) of the total cases. [6]

Air Quality Index (AQI) is a metric to measure or compute how polluted the air is around a particular location. The Air Quality index is used by government firms to check the status of air quality in certain locations and each government firm has its air quality indices. The air quality index is sorted into six brackets i.e Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe. To state an example, in India, if the air quality index of a certain region is above 300, it is categorized in the ”Very Poor” bracket, and this adverse AQI is caused by the pollutants present in the air such as Nitrogen dioxide (NO2), Carbon monoxide (CO), Sulfur dioxide (SO2), Ozone (O3), Benzene, Toluene, Xylene, PM2.5, and PM10.

The main objective of this novel research is to compare the state of the air quality index of India before and after the global pandemic by measuring the pollutants namely Nitrogen dioxide, Sulfur dioxide and Carbon Monoxide along with PM 2.5 in the air at different time intervals. Furthermore, this study focuses on developing forecasting models thorough time-series analysis to accurately predict the daily air quality index.

II. Related Work

Tanisha Madan et.al in their study used machine learning in the prediction of air quality index [7]. The machine learning models that were used in the study were logistic regression, random forest, linear regression , support vector machine (SVM), XG Boost and Hybrid Tree And Light Gradient Boosting Model. It was observed that the accuracy achieved by the Hybrid Tree And Light Gradient Boosting Model method was more than 99% which was more in comparison with other machine learning models. Chavi Srivastava et.al provided a comprehensive analysis of air

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health warnings of emergency conditions. The entire population is more likely to be affected.
Hazardous	301 to 500	Health alert: everyone may experience more serious health effects

Fig. 1. AQI Bracket of India

pollution in Delhi using various machine learning techniques [8]. In the study, various regression and classification techniques were implemented to estimate air quality. The models that were used were support vector machine (SVM), linear regression, decision tree classifier, SDG regression, random forest classifier, gradient boosting regression, adaptive boosting regression and artificial neural network (ANN). The models were trained and tested using the AQICN dataset. The authors stated that support vector regression (SVR) and neural network (MLP) performed better overall than other machine learning techniques.

Nimisha Tomar et.al offered a different procedure to forecast air quality index. In their study, they used auto regression models to predict air quality [9]. In the research, the authors also implemented another forecasting model ARIMA (auto regressive integrated moving average) model. In the research analysis, the forecasting models gave a prosperous outcome and consequently can be used to forecast air quality index value. S Sankar Ganesh et.al conducted a study on forecasting air quality index using regression models on the Delhi and Houston AQI dataset [10]. The authors implemented machine learning models such as (SVR) support vector regression and multiple linear regression which includes mini-batch gradient descent, gradient descent and stochastic gradient descent. The main goal of the study was to compare different regression models to forecast AQI. It was found that the support vector regression (SVR) technique manifested highest accuracy as compared to other regression models.

Jasleen Kaur Sethi et.al proposed a study in which they monitored the impact of air quality on the COVID-19 fatalities in Delhi using various machine learning techniques [11]. The authors used various regression and classification techniques like decision tree, linear regression and random forest to extract the main pollutants that impacted the COVID-19 mortality rate. The result showed that NH3, ozone, NO2 and PM10 were the major pollutants.

Atharav Barve et.al proposed a new method to forecast air quality index using a parallel dense neural network and LSTM cell [12]. The dataset used in the research paper was the air quality index of Beijing. The author's found that the combination of LSTM cells and dense neural networks performed much better than the standard LSTM model. The researchers also found that the model performed more efficiently, as when the duration of time of forecasting increases, the performance of the model will play a major factor.

Dat Q. Duong et.al conducted a study on predicting air quality index using multi-source machine learning [13]. The authors used datasets from various sources SEPHLAMediaEval 2019", MNR-Air-HCM," and MNR-HCM. The models that were used in the research paper were SVM, Random Forest, XGBoost, LightGBM, and CatBoost. In the result, they found that random forest models performed much better than other models.

III. Methodology

During this study of ours, we will be comparing the state of air quality in India, before and after the global coronavirus pandemic. We will also be utilizing a variety of machine learning models and time-series analysis techniques for accurately forecasting and predicting the air quality index of India. The main approach that we are adopting to train our model includes firstly, data splitting where we will be dividing the model into training and testing. Secondly, we preprocess the data using standardization and missing value computation. We used EDA (Exploratory data analysis) for a better air quality analysis during the coronavirus pandemic. Lastly, we compared the results for time series analysis and the traditional machine learning model using various regression metrics.

A. Dataset

The initial step in creating a machine learning model is gathering relevant data in that specific field. Datasets are the stepping stone for training and testing the model and maximizing its accuracy. In this study for air quality analysis and forecasting, the dataset was compiled from the Central Pollution Control Board (CPCB) website [14], which is the official body of the Government of India. The dataset contains hourly and daily levels of air quality and AQI (Air Quality Index) adjusted for the Indian pollution bracket for various stations across multiple Indian cities. The information outline

begins from 2015 to 2020 and consists of 16 columns which include 15 attributes and 1 target column with the desired Air Quality Index.

B. Data Preprocessing

Data preprocessing is a methodology that can be utilized to translate unstructured data to appropriate structured formatted data which can further help in analyzing data. The data input of the air quality index that is collected from the source is unprocessed, it needs processing which can further assist in getting a satisfactory accuracy of the model. There are various types of data preprocessing such as data cleaning, data transformation and data reduction.

In the dataset, there were many missing values in various columns, due to which different machine learning models will fail to produce an output. To avoid the failure of the machine learning model, we used SimpleImputer which provides various basic techniques for substituting missing values. The imputer function transforms the missing data either with a constant value or using statistics (median, mean or most frequent). Here we utilized it to impute missing values with the mean values for each column in which missing values were located. In the paper, we used a mean value strategy in the imputer function to replace the missing value, this process is also known as data refinement. After defining and padding all the values in the dataset, we split the dataset into train data and test data for further preprocessing the data.

In the dataset, there are many different features which have different values of data, so hence it is very necessary to scale these features to a discrete range. The feature scaling makes sure that an attribute which is not much important with a larger range of data can suppress an attribute which is significantly more important. Hence, we used a standardization technique to scale all the values in the dataset. The formula of standardization is given as follows:

$$Z = \frac{(x-\mu)}{\sigma}$$

μ = mean of the given distribution
σ = standard deviation of the given distribution

C. Exploratory Data Analysis (EDA)

In Fig. 3, we have implemented a correlation graph for the air quality index dataset in a heatmap manner. In order to visualise the data in the heatmap, the colour dark green basically shows that the attributes on x axis and y axis correlate with each other strongly and dark pink shows that the attributes are negatively correlated with each other.
In the heatmap, we can visualize that the AQI is positively correlated with PM2.5, PM10, NO, NO2, NOx, CO, SO2 and negatively correlated with NH3, O3, Benzene , Toluene, Xylene. From the heatmap, we can visualize that the PM2.5, PM10, NO, NO2, NOx, CO and SO2 are major factors in the forecasting of the Air quality index.

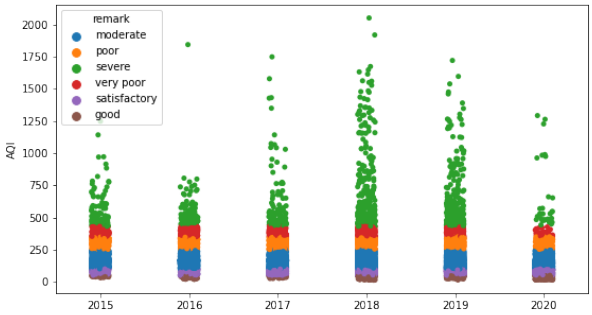

Fig. 2. Correlation Matrix

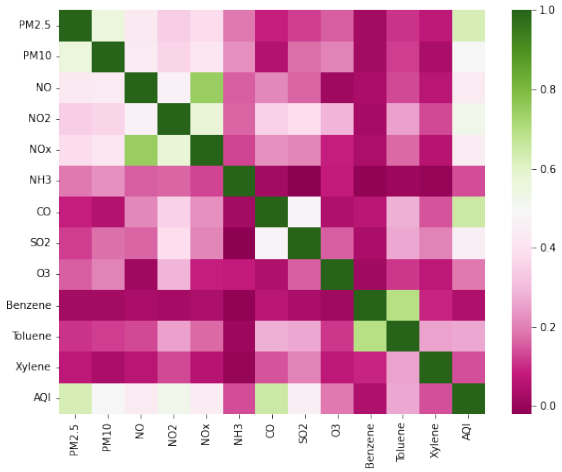

Fig. 3. Correlation Matrix

D. Models

1) Linear Regression: Due to its simple and straightforward representation, Linear Regression is a very popular model which is used extensively in the field of machine learning and statistics. The representation is a linear equation which evaluates a set of input values (x) to predict an output (y). For a simple regression problem with one independent variable, the form of the model would be:

$$y = \theta_1 + \theta_2 x$$

It fits the model with the best linear line to predict the values y for given values of x. We have used this model to predict the air quality index.

2) SGDRegressor: The Stochastic Gradient Descent Regressor (SGDRegressor) is a very effective and popular algorithm for linear models with the benefit of being simple and easy to understand. It is mostly used when the size of data is very large and the number of samples and features is very high. The SGDRegressor is able to provide different loss functions and penalties and is able to fit linear models.

It works by continuously running through the training dataset and when it meets with a training example, the parameters are updated with the gradient of the error with respect to that single training example.

3) XG Boost Regressor: The Extreme Gradient Boost Regressor or XG Regressor is a gradient boosting algorithm. It is a class of ensemble machine learning algorithms used for classification or regression predictive modelling problems. It is an efficient implementation of gradient boosting that can be used for regression predictive modelling.

4) RandomForestRegressor: A Random Forest is an ensemble technique. The ensemble is a way in which we combine predictions from two or more models for the final result. A Random forest is capable of performing both regressions as well as classification. It uses multiple decision trees and bagging. In bagging, different sample data from the dataset is used to train different decision trees where sampling is done with replacement, and then the final decision is combined. This controls the over-fitting and helps in improving the predictive accuracy of the model.

5) SVR: Support Vector Regression or SVR is one of the most popular and widely used algorithms. It is used for classification problems in machine learning. The SVR is a supervised learning model that analyzes data that is used for classification and regression while associating different learning algorithms. The SVM algorithm finds a hyperplane in n-dimension that distinctly classifies the data points. The data points that are closest to this hyperplane or on it are called support vectors. The SVR tries to fit the best line within a value that is between the hyperplane and boundary line. Let's assume that the hyperplane equation is:

$$Y = wx + b$$

Then the equations of decision boundary become:

$$wx + b = +a$$
$$wx + b = -a$$

Thus, the hyperplane satisfies $-a < Y - wx + b < +a$ to meet SVR criteria.

6) Sarima: A Seasonal auto-regressive integrated moving average or Sarima is a time series model based on seasonal trends. A time series is an order or list where the data is recorded over a regular time interval. The interval can be of yearly, quarterly, monthly, weekly, daily, hourly, minutes or even seconds. As we know that Arima models are good with data that are in trend but it does not support seasonal components. For this reason, Sarima is introduced. It is an extension of Armia that supports uni-variate time series data, auto-regressive and moving average elements. It has three new hyper-parameters, the auto-regression(AR), differencing(I), and moving average(MA) for the seasonal component and some additional parameters for the period of seasonality.

IV. Result

The daily Air Quality Index was monitored and predicted using the dataset made available by the Central Pollution Control Board for all four seasons during the period of 2015 to 2020. In the research, to forecast AQI in India we implemented non-time series models and time series models. For non-time models, we implemented Linear Regression, SGDRegressor, XG Boost Regressor, RandomForestRegressor and for the time series model, we implemented the Sarima model. We also applied hyper-parameter tuning to the models in order to increase performance and evaluated various evaluation metrics such as R^2, RMSE and MAE to assess and compare the performance of each model.

The R^2 score and other evaluation metrics were calculated after hyper-parameter tuning using mentioned machine learning models. The highest score was achieved in the case of the Random Forest Classifier and Sarima Models both with a score of 0.879. XGBoost Regressor showed an R^2 score of 0.85 while the Linear Regression model and SGD Regressor model showed the score of 0.79 and 0.78 respectively. Lastly, in the case of the support vector regression model, the R^2 score comes out to be 0.628.

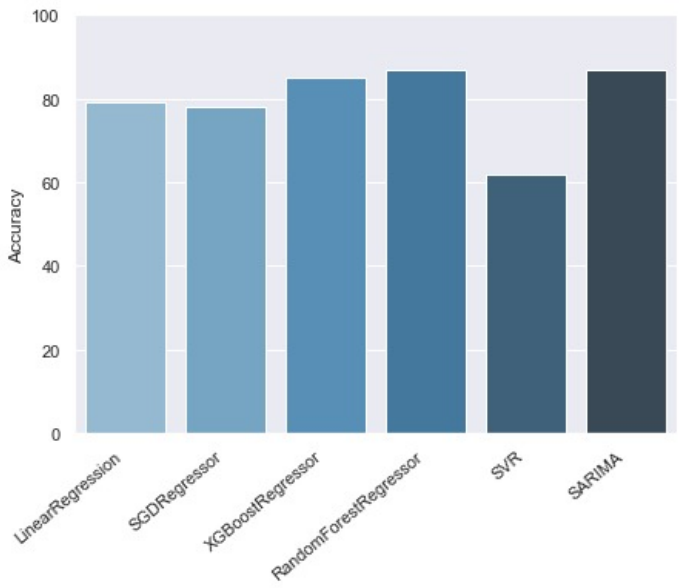

Fig. 4. Accuracy Bar Plot

A smooth distribution line was also plotted for Random Forest Regressor by a seaborn library Distplot to show the performance of the model.

The Fig. 6 shows the integrity of fit with the forecasts imagined as a line. A decent fitting model was observed to foresee the air quality index based on the region specific to India.

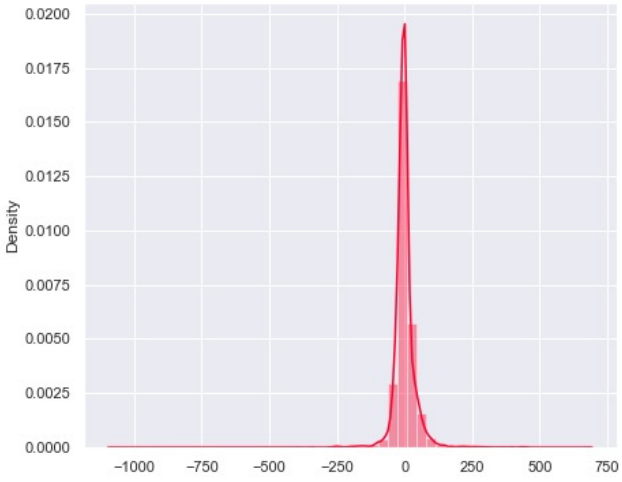

Fig. 5. Distplot for Random Forest Regressor

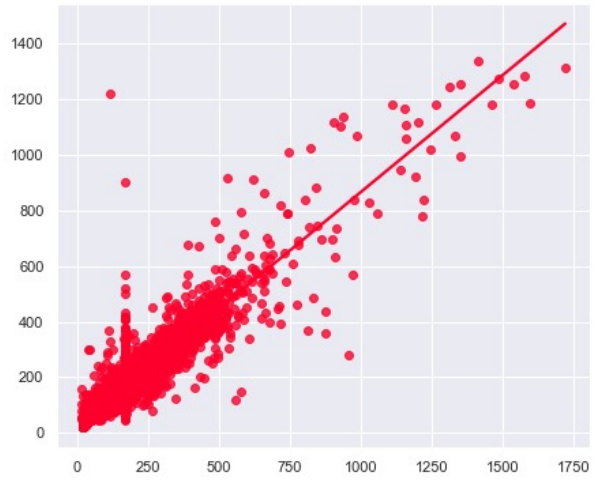

Fig. 6. Scatter Plot for Random Forest Regressor

V. Conclusion

From this study, we deducted that the air quality index was significantly reduced to minimal levels during the lockdown forced by the Government. The responsible reasons for this note-worthy drop included a reduction in the number of vehicles on the roadway and the shutting down of multiple non-essential industries and factories.

The present study also focuses on forecasting the daily air quality index by methods including time-series analysis techniques and numerous machine learning models. Out of the experimented models, we concluded both the time-series model Sarima as well as Random Forest Classifier provided decent results with the R^2 score of 0.879

References

[1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, and W. Tan, "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, 2020, pMID: 31978945.

[2] I. R.J., "The dramatic impact of coronavirus outbreak on air quality: Has it saved as much as it has killed so far?" in *Global Journal of Environmental Science and Management*, 2020.

[3] R. D. Brook, S. Rajagopalan, C. A. Pope, J. R. Brook, A. Bhatnagar, A. V. Diez-Roux, F. Holguin, Y. Hong, R. V. Luepker, M. A. Mittleman, A. Peters, D. Siscovick, S. C. Smith, L. Whitsel, and J. D. Kaufman, "Particulate matter air pollution and cardiovascular disease," *Circulation*, 2010.

[4] G. B. Hamra, N. Guha, A. Cohen, F. Laden, O. Raaschou-Nielsen, J. M. Samet, P. Vineis, F. Forastiere, P. Saldiva, T. Yorifuji, and D. Loomis, "Outdoor particulate matter exposure and lung cancer: A systematic review and meta-analysis," *Environmental Health Perspectives*, 2014.

[5] T. J. Woodruff, J. D. Parker, and K. C. Schoendorf, "Fine particulate matter (pm¡sub¿2.5¡/sub¿) air pollution and selected causes of postneonatal infant mortality in california," *Environmental Health Perspectives*, 2006.

[6] M. S. S. M. C. O. F. P. H. M. H. F. J. P.-T. V. Q. P. S. J. S. R. V. J. S. H. Künzli N, Kaiser R, "Public-health impact of outdoor and traffic-related air pollution: a european assessment," in *Lancet*, 2000.

[7] Shrddha and D. Virmani, "Air quality prediction using machine learning algorithms –a review," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020.

[8] C. Srivastava, S. Singh, and A. P. Singh, "Estimation of air pollution in delhi using machine learning techniques," in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, 2018.

[9] N. Tomar, D. Patel, and A. Jain, "Air quality index forecasting using auto-regression models," in *2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, 2020.

[10] S. S. Ganesh, S. H. Modali, S. R. Palreddy, and P. Arulmozhivarman, "Forecasting air quality index using regression models: A case study on delhi and houston," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017.

[11] J. K. Sethi and M. Mittal, "Monitoring the impact of air quality on the COVID-19 fatalities in delhi, india: Using machine learning techniques," *Disaster Medicine and Public Health Preparedness*, 2020.

[12] A. Barve, V. Mohan Singh, S. Shrirao, and M. Bedekar, "Air quality index forecasting using parallel dense neural network and lstm cell," in *2020 International Conference for Emerging Technology (INCET)*, 2020.

[13] D. Q. Duong, Q. M. Le, T.-L. Nguyen-Tai, D. Bo, D. Nguyen, M.-S. Dao, and B. T. Nguyen, "Multi-source machine learning for aqi estimation," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020.

[14] "Central pollution control board."