IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 7, July 2022

DOI: 10.17148/IJIREEICE.2022.10743

AIR QUALITY INDEX ANALYSIS USING MACHINE LEARNING

Jeevan. S[1], H.L Shilpa[2]

PG Scholar, Dept. of MCA, PES College of Engineering, Mandya, Karnataka, India[1]

Assistant Professor, Dept. of MCA, PES College of Engineering, Mandya, Karnataka, India[2]

Abstract: Air pollution is an important environmental risk factor in propagation diseases such as lung cancer, autism, asthma and low birth weight etc. Regulation of air quality is an important task of the government in developing countries for ensuring people’s health and welfare.
Air pollution differs from place to place and depends on multiple pollutant sources such as industrial emissions, heavy traffic congestions, temperature, pressure, wind, humidity and burning of fossil fuels etc.
Analyzing and protecting air quality has become one of the most required activities for the government in almost all the industrial and urban areas today. In this paper, machine learning algorithms are used to analyze the concentrations of air pollutants such as SO_2, NO, $PM_{2.5}$, O_3 and PM_{10}.
This model analyses the air quality based on various pollutant concentrations through visualizations for effective feature extraction and decision making. A machine learning model is built using linear regression and SARIMA model to predict the air quality index based on past air quality data. The experimental results show that the proposed model can be efficiently used to detect the quality of air and predict the level of air quality in the future. The model has scored 71.69% for the train data.

Keywords: Air Pollution, air pollutants, air quality index.

I.INTRODUCTION

Due to new invention, development is accelerating quickly. This, combined with a rapid increase in the number of people and vehicles, will result in a number of serious environmental issues, including air pollution, sound pollution, deforestation, water pollution, acid rain, the release of harmful materials, and others. To meet the demands of an expanding population, there has been a dramatic growth in industrialization, which could result in the release of hazardous gases into the atmosphere from a variety of businesses, creating a severe problem with air pollution in city areas around the world. This indicates that the air people are breathing is not clean but rather polluted due to the abundance of dangerous gases and other airborne particles that have a negative impact on people's health. Pollution causes a decline in air quality.

In most of the city areas the pollution of air becomes a severe worry. The citizens ought to be aware of the air they are breathing. The National Ambient Air Monitoring Network produces data that shows the concentration level of different air contaminants, however this data is difficult for the average person to understand. As a result, India's cities national Air Quality Index (AQI) is created by the Central Pollution Control Board (CPCB). The air quality index (AQI) provides information on the level of air pollution in a certain area. This indicates that the AQI measures the actual air quality around us in a manner that is qualitative and associated with different negative health effects.

According to CPCB, the AQI is going to be calculated using 12 parameters (Air Pollutants) including NO_2 (Nitrogen Dioxide), SO_2 (Sulfur Dioxide), CO (Carbon Monoxide), O_3 (Ozone), PM_{10} (Particulate Matter having diameter 10 micron or less), $PM_{2.5}$ (Particulate Matter having diameter 2.5 micron or less), NH_3 (Ammonia), C_7H_8 (Toluene), C_8H_{10} (X (Benzene). Most of the time, the AQI is based on the criterion pollutants (i.e., PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, and O_3), however it’s difficult to use many pollutants from the list of 12 pollutants when calculating the AQI. The extraction of pollutants, however, is based on the AQI objectives, the period for averaging, the data that is available, the frequency of monitoring, and the measurement techniques. A simple definition of AQI is that it is a number that is used by government organizations to gauge air pollution levels and inform the public. A significant portion of the population will be impacted if the AQI rises since it will have a hazardous impact on human health. As is well known, the concentration of various air pollutants can be used to calculate the AQI, which results in a single numerical number known as the AQI.

II.LITERATURE SURVEY

J. Rene Beulah and K. Mahesh Babu [1] The analytical process started with data preparation and cleansing, missing record detection, comprehensive evaluation, and model construction and evaluation. When comparing accuracy with classification record, the decision tree technique procedure has excellent accuracy on the public test set. This model can assist India's urban areas in forecasting the future of air quality and it's reputation, depending on their capacity to act.

Mohammad Ashraf Ottom and Khalid M O Nahar [2] The identification of the Air Quality Index (AQI) in the many nations, each of which employs a distinct management strategy, is necessary for assessing the quantity of pollutants and pollution. The most polluting factor that contributes to the impurity of the air is specified in the classification of the AQ as a polluted area. This model employs machine learning algorithms like Logistic Regression and Decision Trees (DT). The most polluting component may be predicted using this model and daily observations.

Shivam Pisal, Ritik Sharma, and Gaurav Shilimkar [3] This model used Jupiter notebook to implement various machine learning algorithms in Python. The model starts by choosing the essential features that are taken into account for the result are correlated and can thus be taken into account to train the model. By using unique calculations like direct relapse, Decision Tree, and SARIMA Model, this model forecasts the air quality list. From the findings, it is clear that the SARIMA Model calculation provides a more accurate prediction of air quality than the Decision Tree.

Laxmi Chaudhary and Avnish Bora [4] The advanced IOT devices with a variety of sensors and devices are used by this model to collect the essential data for model training. The air quality characteristics essential for the future development of cities can be analyzed and measured using machine learning-based AI algorithm models. These new technologies, which feature compact devices and inexpensive sensors, effectively provide vast amounts of data to the model. It contributes to lowering pollution of air, which lowers the risk of health problems for people and also prolongs the lives of other organisms, including plants and animals, in our environment.

Heniel Kashyap and Udit Ranjan Kalita [5] The primary challenge that many emerging nations deal with is air pollution. These days, pollution is getting worse and worse every day for a variety of reasons, including industry, population increase, the development of new technologies, the chemical industry, and many other sectors. This approach uses semiconductor sensors like the MQ9 and MQ7, which can identify some factors that contribute to air pollution and forecast the AQI.

Zeynep Cansu Ayturan and Yasin Akn Ayturan [6] Deep learning-based air pollution modelling is a novel idea. This programme anticipates future air pollution and uses deep learning to produce findings that are remarkably accurate. Deep learning algorithms and methods may help to reduce the lack of relevance in air pollution predictions. Deep learning can be used to predict the AQI value from a variety of unstructured data formats, including images, sounds, and numerical data. Some examples of techniques used for this are LSTM and CNN. Their effectiveness mostly depends on the algorithms used and the cleansed data.

R Soria, S. Berres, B. Mark, and L. Caro [7] We can infer from this study that an auto encoder neural network can be successfully used to detect air contaminants using sensors. This model demonstrates a high level of normal pattern identification accuracy. The likelihood of recognising the ambiguous data, however, has marginally increased in the detection of air contaminants. Such minute differences may cause the model's accuracy to decline. The ideal configuration is model that, in the worst scenario, when test into a dataset with an accuracy of 80%, is able to detect 50% of air pollutants, which is considered to be useful in real-world application scenarios.

Rogulski, M., and Badyda, A. [8] From the standpoint of metropolitan regions, the creation of a new AQI monitoring paradigm is crucial. According to research, real-time air pollution data can assist inform the public about the best preventative measures based on their unique health needs and increase awareness of air pollution. The AQI is predicted by this model using data cleaned from multiple environmental measurements and input from numerous sensors.

Tomasz Rymarczyk and Tomasz Cieplak [9] Due to the impact on human health, measuring the contaminants that create air pollution is a subject that is being examined more quickly. As a result, it appears important to develop algorithms that can use publicly available datasets to track the quality of the air in various regions. Data cleaning techniques based on machine learning are available. The loop method was used to create the model that contained the analysis of outlier outcomes for various air contaminants. In addition to making results simple to understand, the applied method is helpful in the analysis of data streams. The presumed quality of the input data can be maintained by using outlier detection techniques.

Nikolaos Avouris and Elias Kalapanidas [10] We can determine that three algorithms are correctly forecasting the approximate value by contrasting the outcomes from different models. This model demonstrates that CART classifiers

DOI: 10.17148/IJIREEICE.2022.10743

outperform other tree-induction classifiers in complicated and multidimensional air quality prediction data. All of the algorithms' results are quantitative, yet they might nevertheless show a discrete choice. The algorithms' total performance, however, can be characterized as noteworthy given that the human being specialists at the Greek AQOC do not perform more accurately than these levels. The findings suggest that management and prediction of air quality can benefit from the application of machine learning techniques by academics and practitioners.

III. PROPOSED SYSTEM

The proposed model is designed with the intention that, common people need to be aware of Air Quality Index (AQI), it's a metric that determines the quality of the air. Whether the air is hazardous for human and other living organisms in the environment.

The model begins with collecting the necessary data required for building the model. The dataset is taken from Central Pollution Control Board (CPCB) website. After gathering the required data need to pre-process the collected data. Feature selection takes place that means un-necessary records are dropped, and necessary data to predict the AQI is taken into consideration. And visualization of dataset is appeared.

Then data cleaning procedure taken place, this needs to be done to get more accuracy of the model. Under this phase missing value treatment, checking the outlier are taken place to increase the score of the model. Later Machine Learning algorithms such as linear regression & SARIMA Model are applied to the cleaned train data to train the model and then test data is used to predict the accuracy score of the model. The model gives the AQI value and also alerts about the category under which the AQI value is lie down.

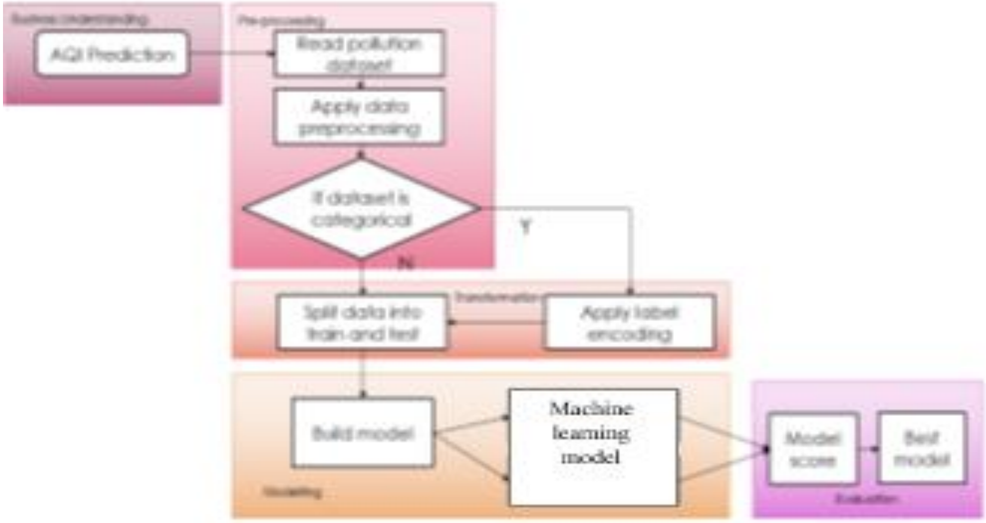


Figure 1: Flowchart of the proposed model

Then data cleaning procedure taken place, this needs to be done to get more accuracy of the model. Under this phase missing value treatment, checking the outlier are taken place to increase the accuracy of the model. Later Machine Learning algorithms such as linear regression & SARIMA Model are applied to the cleaned train data to train the model and test data is used to predict the accuracy of the model.

IMPLEMENTATION AND PROCESS

A. Dataset

The dataset available from CPCB website contains approximate total of 30,000 data tuples with over 16 attributes. The high dimensionality had an adverse impact on the model building. The attributes included pollution level prediction for 4 different pollutants. Thus, we trimmed down to one pollutant - NO_2. The same model can be applied to different pollutant attributes as per requirement. Also, the data provided is for 6 years with hourly values for all the days.Dataset included different parameters such as Particulate Matter ($PM_{2.5}$ and PM_{10}), Nitrogen monoxide (NO), Nitrogen Dioxide (NO_2), Nitrogen Oxide (NOx), Ammonia (NH_3), Carbon Monoxide (CO) Sulphur Dioxide (SO_2), Ozone (O_3).

B. Data Pre-Processing

For data analysis, plot most polluted cities based on AQI, PM_{10} and CO parameters. Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. The following diagram shows the most polluted cities presented in the dataset based on the AQI, PM_{10} and CO parameters value.

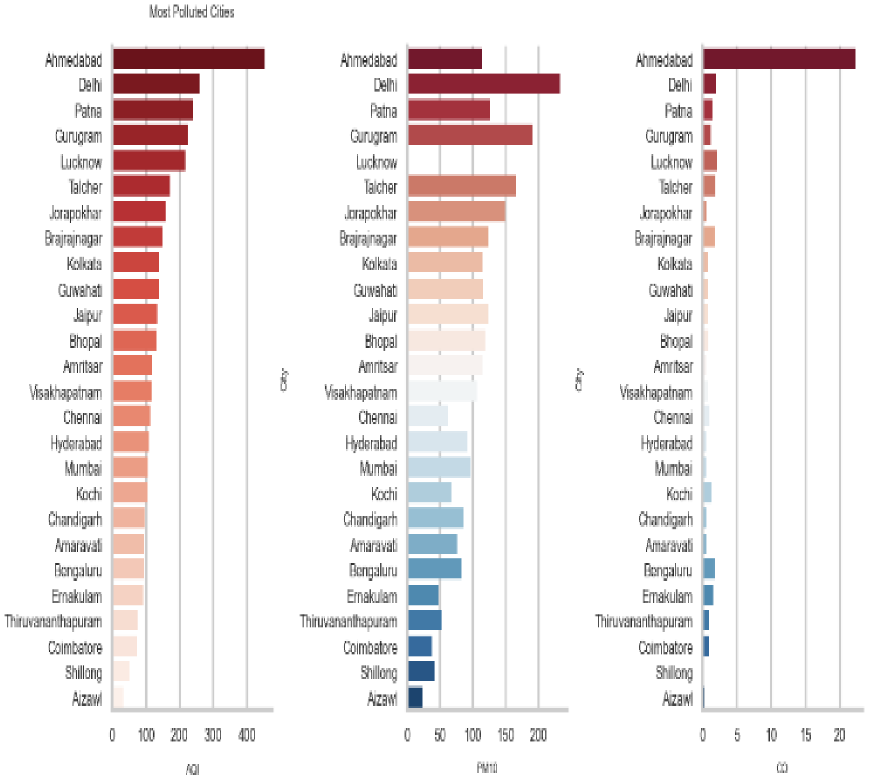


Figure 2: Most polluted cities

C. Attribute Selection

The new attribute is selected from the given set of attributes. The attributes which majorly contribute to air pollution and the row-wise highest value is considered as Air Quality Index.

D. Standard Scaler

In Machine Learning, StandardScaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1. In this model, I will walk you through how to use StandardScaler in Machine Learning. StandardScaler is an important technique that is mainly performed as a preprocessing step before many machine learning models, in order to standardize the range of functionality of the input dataset.

StandardScaler comes into play when the characteristics of the input dataset differ greatly between their ranges, or simply when they are measured in different units of measure.

StandardScaler removes the mean and scales the data to the unit variance. However, outliers have an influence when calculating the empirical mean and standard deviation, which narrows the range of characteristic values.

These differences in the initial features can cause problems for many machine learning models. For example, for models based on the calculation of distance, if one of the features has a wide range of values, the distance will be governed by that particular characteristic.

The idea behind the StandardScaler is that variables that are measured at different scales do not contribute equally to the fit of the model and the learning function of the model and could end up creating a bias.

So, to deal with this potential problem, we need to standardize the data ($\mu = 0$, $\sigma = 1$) that is typically used before we integrate it into the machine learning model.

Standardize features by removing the mean and scaling to unit variance.

The standard score of a sample x is calculated as:

$z = (x - u) / s$

DOI: 10.17148/IJIREEICE.2022.10743

E. Data Cleaning

Data cleaning is an important part of the data preparation process. If we don't clean the raw data, the model's results may suffer, and the model may not suit the dataset. Cleaning data manually is insufficient; thus, I use the procedures below to clean data: For improved results, check for missing numbers and fill all null values with the median function. Check for duplicate values and eliminate them.

F. Data Visualization

Data visualization is the graphical representation of information and data and it plays an important role in the portrayal of both small-scale and large-scale data. Graphical elements like charts, graphs, and maps, data visualization tools provide an approachable way to see and fathom trends, outliers, and patterns in data.

G. Linear Regression

Linear regression is probably the method where most of the academicians started their first machine learning experience. Its main working principle lies behind the fitting of one or more.

Independent variables with the dependent variable into a line in n dimensions n usually denotes the number of variables within a dataset. This line is supposedly created as it would be minimizing the total errors when trying to fit all the instances into the line. Under machine learning, linear regression is equipped with the capability to learn continuously by optimizing the parameters in the model.

These parameters are including x_0, x_1, x_2,…., x_p. Most commonly, optimization is carried out by a method called gradient descent. It works by partially deriving the loss function and all parameters will be updated by subtracting the previous value with the derivative times a specified learning rate.

The learning rate can be tuned by the simplest way, which is rule of thumb (trial and error), or a more sophisticated rule, e.g., meta-heuristic. Another parameter that is left for tuning is the amount of generalization added to the model. Regularization is undergone as an art to lessen the chance of overfitting and increase the robustness of the model.

Two types of regularization used in linear regression are lasso and ridge regression. Lasso regularization will eliminate less important feature by letting the feature's coefficient to zero, and retain another more important one. Ridge regularization on the other hand will not try to eliminate a feature, but instead, tries to shrink the magnitude of coefficients to get a lower variance in the model.

H. SARIMA Model

SARIMA is Seasonal ARIMA, or simply put, ARIMA with a seasonal component. As mentioned above, ARIMA is a statistical analysis model that uses time-series data to either better understand the data set or to predict future trends.

Modeling a time series data is a highly subjective and individual process. One may have different parameters for the same time series. Hence, there is no fixed solution. The best solution is the one that successfully fulfills the business requirements. Owing to this level of subjectivity involved, it sometimes gets tough to understand the model building process.

Several studies, tutorials, and implementations later, I was able to crunch the findings into a framework.

The mean and the variance of the data remains same throughout the data. Hence, there is no need to transform the data. We now proceed to check the trend and seasonal components of the data.

Once we have a fitted model to the data, it is necessary to check the residual plots to verify the validity of the model fit. A good forecasting method will yield residuals with the following properties:

The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals that should be used in computing forecasts.The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.

IV.EXPERIMENTAL RESULTS

This model is implemented using traditional machine learning techniques such as linear regression and SARIMA Model in python PyCharm Integrated Development Environment (IDE). Since the air quality index value is dependent on the amount of parameters present in the air. If the value of the independent variables changes then value of the feature also changes, so we can conclude that features are correlated with the independent variables.

The performance of the model can be measured by different evaluation metrics such as Mean Square Error (MSE), R-Square error and root mean square error (RMSE). For the experiments, the data is split into train and test data, the model considering 80% for training and 20% for testing which is used to check whether the model is working properly or not.

DOI: 10.17148/IJIREEICE.2022.10743

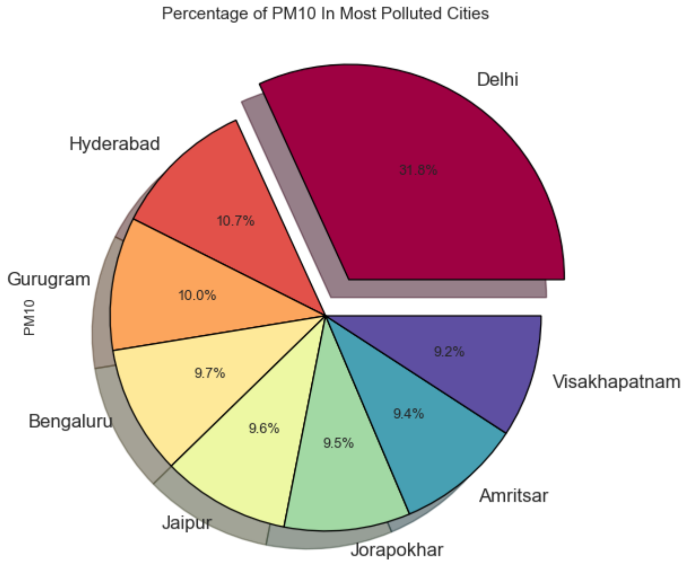


Fig 3: PM_{10} presence in most polluted cities

The above figure depicts that presence of PM_{10} pollutant in most polluted cities which are present in the dataset. Delhi has the most percentage of presence of PM_{10} that is 31.8% and Vishakhapatnam has the least percentage of presence of PM_{10} that is 9.2%.

The following figure depicts that presence of CO pollutant in most polluted cities which are present in the dataset. Ahmedabad has the most percentage of presence of CO that is 60.8% and Talcher has the least percentage of presence of CO that is 2.8%.

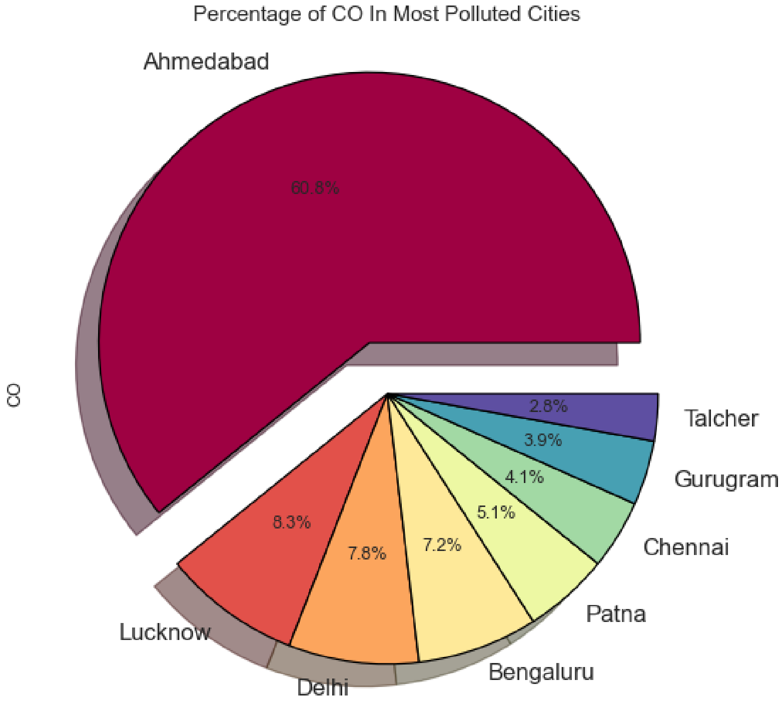


Fig 4: CO presence in most polluted cities

DOI: 10.17148/IJIREEICE.2022.10743

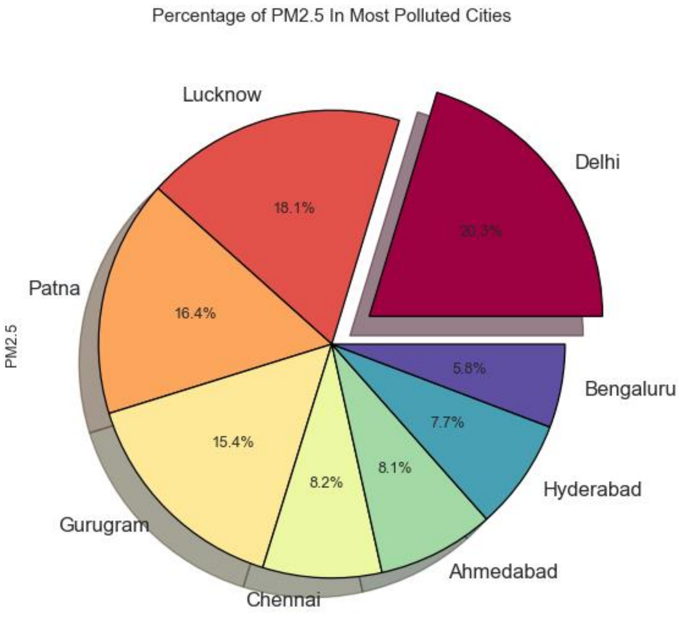
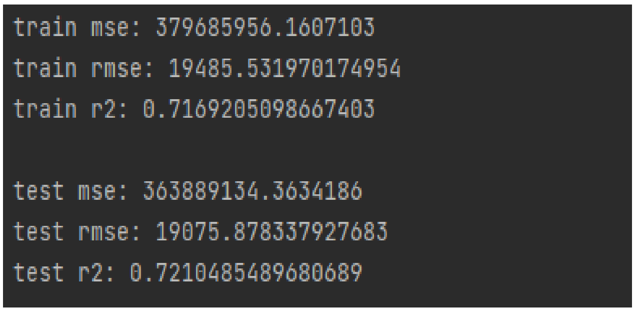


Fig 5: $PM_{2.5}$ presence in most polluted cities

The above figure depicts that presence of $PM_{2.5}$ pollutant in most polluted cities which are present in the dataset. Delhi has the most percentage of presence of $PM_{2.5}$ that is 20.3% and Bangalore has the least percentage of presence of $PM_{2.5}$ that is5.8%.



```
train mse: 379685956.1607103
train rmse: 19485.531970174954
train r2: 0.7169205098667403

test mse: 363889134.3634186
test rmse: 19075.878337927683
test r2: 0.7210485489680689
```

Fig 6: Accuracy of the model

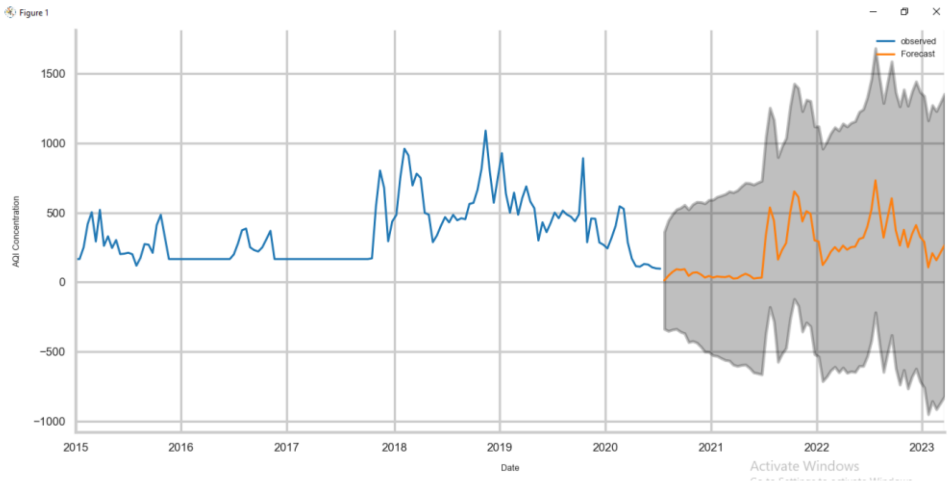


Fig 7: Future Prediction of AQI

DOI: 10.17148/IJIREEICE.2022.10743

The above figure depicts the future prediction of air quality index value based on the past data present in the dataset we had taken. The blue plot represents the past data which is already present in the dataset and the orange plot represents the future predicted AQI value.

V CONCLUSION

The purpose of this project is to know in detail about the Air Quality Index (AQI) as AQI tells whether the air around us is polluted or not. It is important to know about AQI because unless and until the people know the worst impacts or hazards of air pollution they will not become that much aware about the air pollution and try to reduce it. This application can help India meteorological division in predicting the way forward for air nice and its reputation and will depend on that they are able to take motion. R-square is the evaluation metric that is used to check the model accuracy that is 71.69% for train data and 72.10%. Additionally, it is expected to continue with modules like determining which pollutant has triggered the value, which age category people are affecting more and what are the precautionary methods need to be taken care and what are the prevention methods to reduce the AQI needs to be display on the model.

VI REFERENCES

TEXT BOOKS

- Geron Aurelien. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd *Edition*, Kindle *Edition*, O'Reilly Media, 13 March 2017.
- **Gary B. Shelly. Systems Analysis and Design, Shelly Cashman Series' textbooks, 1991.**
- Oliver Theobald. Machine Learning for Absolute Beginners: A Plain English, Independently Published, 2017.

WEBSITES

- https://www.javatpoint.com/linear-regression-in-machine-learning
- https://www.javatpoint.com/machine-learning-random-forest-algorithm
- https://www.youtube.com/watch?v=nxFG5xdpDto

RESEARCH PAPERS

[1] K. Mahesh Babu, J. Rene Beulah. Air Quality Prediction based on Supervised Machine Learning Methods. International Journal of Innovative Technology and Exploring Engineering (IJITEE), July 2019.

[2] Khalid M O Nahar, Mohammad Ashraf Ottom, Fayha Alshibli and Mohammed M. Abu Shquier. AIR QUALITY INDEX USING MACHINE LEARNING. COMPUSOFT, An international journal of advanced computer technology, September-2020.

[3] Ritik Sharma, Gaurav Shilimkar, Shivam Pisal. Air Quality Prediction by Machine Learning. International Journal of Scientific Research in Science and Technology, May-June-2021.

[4] Avnish Bora, Laxmi Chaudhary. Technological Advancements in Air Pollution Monitoring Systems. International Journal of Engineering Research and Management (IJERM), November 2020.

[5] Udit Ranjan Kalita, Heniel Kashyap, Amir Chetri and Jesif Ahmed. Centralized Air Pollution Detection and Monitoring. ADBU Journal of Electrical and Electronics Engineering (AJEEE), February 2018.

[6] Yasin Akın Ayturan, Zeynep Cansu Ayturan and Hüseyin Oktay Altun. Air Pollution Modelling with Deep Learning. Int. J. of Environmental Pollution & Environmental Modelling, September 20, 2018.

[7] B Mark, R Soria, S Berres, L Caro , A Mellado and N Schiappacasse. Detection of Anomalous Pollution Sensors Using Deep Learning Strategies. IOP Conf. Series: Earth and Environmental Science, June 06, 2020.

[8] M Rogulski and A Badyda. Current trends in network based air quality monitoring systems. 2nd International Conference on the Sustainable Energy and Environmental Development. IOP Publishing, 2019.

[9] Tomasz Cieplak, Tomasz Rymarczyk and Robert Tomaszewski. A concept of the air quality monitoring system with machine learning methods to detect data outliers. MATEC Web of Conferences 252, 2019.

[10] Elias Kalapanidas and Nikolaos Avouris. Applying Machine Learning Techniques in Air Quality Prediction. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2018.