Air Quality Index Prediction using Linear Regression

Ambika G. N, Bhanu Pratap Singh, Bhavya Sah, Dishi Tiwari

Abstract*- controlling and preserving the better air excellence has become one of the most indispensible events in numerous manufacturing plus metropolitan regions at present. The excellence of air is harmfully affecting payable to the various forms of contamination affected via the transportation, power, fuels expenditures, etc. The installation of destructive fumes is spawning the severe hazard for the class of natural life in developed metropolises. Through cumulative air contamination, we require implementing competent air excellence monitoring models which gathers the statistics about the absorption of air impurities and be responsible for calculation of air contamination in each zone. Hence, air excellence estimation plus calculation has come to be a significant study area. The superiority of air is exaggerated by multi-dimensional influences comprising place, time plus indeterminate parameters. The intention of this development is to examine the machine learning based methods for air quality prediction.*

Keywords*- Air Quality Index, Linear Regression, Auto ARIMA Model, Stepwise Regression, Python, Jupyter Notebook, Tableau Public.*

I. INTRODUCTION

Air is any of the greatest critical expected possessions for the endurance plus subsistence for the complete life on this globe. All forms of the life comprising plant life plus wildlife depending on air for their basic endurance. Thus, all breathing creatures require better excellence of air which is free of dangerous fumes for continuing their lives. The increasing populace, vehicles and productions are poisoning all the air at an alarming rate. Dissimilar elements are related with metropolitan air contamination. In order to estimate the air impurity, contaminant constraints are considered in the lower altitudes of the troposphere, which are meticulous. Air excellence sensor devices extent the attentions of particles that have an anthropogenic source and create the hazardous effects during or after the gulp of air by human being. Particles like PM2.5, CO, NO2, NO etc. disturbs the quality of air. Automobiles release enormous amounts of nitrogen oxides, carbon oxides, hydrocarbons and particulates when burning petrol and diesel. Therefore, we want to consider the added Parameters like number of vehicles along with the Pollutants data for prediction of pollution in specific regions of Bangalore.

Air excellence estimation is an imperative technique to monitor and control the air pollution. It aids to identify the parameters that affect the air quality index and provides the user with appropriate safeguards to maintain the air quality index.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author

Mrs. Ambika G. N, Assistant Professor, Department of CSE, BMS Institute of Technology, Bangalore, (Karnataka), India.

Mr. Bhanu Pratap Singh, CSE, BMS Department of CSE, BMS Institute of Technology, Bangalore, (Karnataka), India.

Ms. Bhavya Sah, CSE, BMS Department of CSE, BMS Institute of Technology, Bangalore, (Karnataka), India.

Ms. Dishi Tiwari, CSE, BMS Department of CSE, BMS Institute of Technology, Bangalore, (Karnataka), India.

II. DESIGN

A. System Design:

This project has been categorized into different steps. Each step has a different component or module responsible for one or more tasks to be accomplished or implemented. These steps occur in sequence and in synchronization to predict the AQI of a given region with the highest possible precision. Input to the system by the user will by the selection of the region and month on the map and the output will be the predicted AQI of that month.

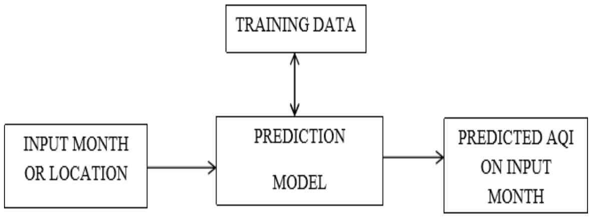


Fig 1: High Level Design of System

B. Component Design:

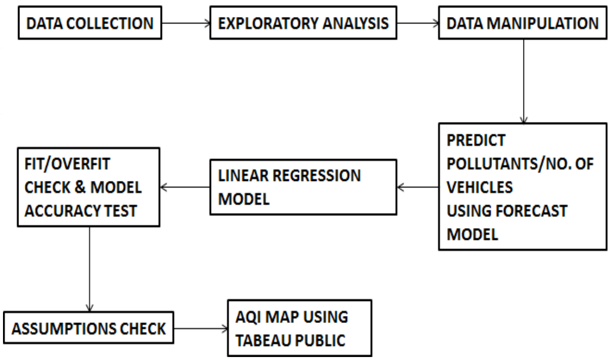

DATA COLLECTION
EXPLORATORY ANALYSIS
DATA MANIPULATION
PREDICT POLLUTANTS/NO. OF VEHICLES USING FORECAST MODEL
LINEAR REGRESSION MODEL
FIT/OVERFIT CHECK & MODEL ACCURACY TEST
ASSUMPTIONS CHECK
AQI MAP USING TABEAU PUBLIC

Figure 2: Flow Chart

The overall process consists of different steps:

1) ***Data collection**:* Data has been collected from different trustworthy sources such as the official government website of Karnataka government.

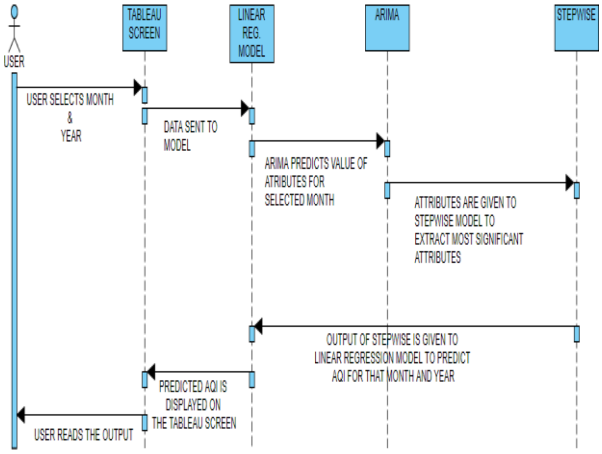

USER
TABLEAU SCREEN
LINEAR REG. MODEL
ARIMA
STEPWISE
USER SELECTS MONTH & YEAR
DATA SENT TO MODEL
ARIMA PREDICTS VALUE OF ATRIBUTES FOR SELECTED MONTH
ATTRIBUTES ARE GIVEN TO STEPWISE MODEL TO EXTRACT MOST SIGNIFICANT ATTRIBUTES
OUTPUT OF STEPWISE IS GIVEN TO LINEAR REGRESSION MODEL TO PREDICT AQI FOR THAT MONTH AND YEAR
PREDICTED AQI IS DISPLAYED ON THE TABLEAU SCREEN
USER READS THE OUTPUT

Figure 3: Sequence Diagram of Events

2) ***Exploratory analysis**:* The identification of outliners, missing values, consistency check etc. all took place in this phase of the project.
3) ***Data manipulation**:* During the Data manipulation phase the missing values were filled in using the mean values of that attribute of data.
4) ***Prediction of parameters using forecast model**:* For linear regression to work properly we need the future values for different parameters such as the levels of NO2, NH2 etc. All these parameters are predicted in this phase of the project.
5) ***Implementation of linear regression**:* As soon as all the parameters become available the linear regression algorithm can be utilized to predict the AQI.
6) ***Fit/Over fit check and accuracy analysis**:* Different sort of errors such as root mean square errors and absolute error percentage are treated as a factor to find out the accuracy as well as to check whether the model is fit or over fit over the training data.
7) ***Assumptions check**:* Linear regression has some assumptions to be checked along with it. All those assumptions are checked during this phase of the project.
8) ***AQI map using Tableau public**:* Project is hosted on the Tableau and made publicly available in this phase. This is the final phase of the project which provides a better view or user interface for user to the project.

C. Sequence Diagram:

As soon as the user enters the month and year the data is sent back. The ARIMA predicts values of different parameters for that month and send them to the stepwise for finding out the most significant attributes which are then send to the linear regression model for predicting the AQI. This AQI is then displayed back on the screen for the user to read.

III. IMPLEMENTATION

A. Dataset:

The first stage of our project is data collection. We have collected data from two government websites:

1. Pollutants data is collected from Karnataka state pollution control board portal: https://www.kspcb.gov.in/ambient_air_quality.html.

2. Data for number of vehicles registered is collected from the given portal: http://transport.karnataka.gov.in/index.php/information/details/vehicle_statistics.

Data fields and their description:

There are 19 fields in the collected data:

1. ***AQI**:* The Air Quality Index (AQI) is a catalog for recording the everyday air excellence. It states us that how hygienic or harmful the air is, and what are the connected wellbeing effects might be a worry.
2. ***STATION**:* The data consist of pollutants and registered vehicle statistics for 20 different stations of Bangalore.
3. ***DATE:*** The collected data includes monthly average data for different regions of Bangalore for 32 months from April-2016 to November-2018.
4. ***ZONE**:* The stations are classified into different zone namely east, west, north, south and yelahanka.

POLLUTANTS:

5. ***PM10**:* The granular particles are between 2.5 and 10 micrometers. These comprises of smolder, filth and dust from manufacturing works, farming and roads as well as mold, spores and pollen.
6. ***PM2.5**:* These elements are 2.5 micrometers or less in diameter. These particles are very minor they can be spotted only with an electron optical microscope. Key sources of the grain particles include automotive vehicles, power plants, uptown wood burning, forest fires, farming burning, etc.
7. ***CO**:* Carbon monoxide is an odorless, colorless gas. It is produced when the carbon in fuels does not entirely burn. Main causes of CO include Vehicle exhaust, fuel combustion etc.
8. ***NH3**:* Gaseous ammonia (NH3) is the greatest abundant alkaline gas present in the atmosphere. The largest source of NH3 emissions is agriculture.
9. ***SO2**:* Sulfur dioxide, a colorless, irritable gas which is created when sulfur-comprising fuels such as coal and oil are burned. Generally, the higher levels of Sulfur dioxide are originated near the enormous industrial complexes. Major sources include power plants, refineries, and industrial boilers.
10. ***NO2**:* Nitrogen dioxide is a part of a cluster of gaseous air contaminants, generated as a result of road traffic and other vestige fuel combustion processes.

11. *PB*: Lead is a naturally occurring toxic metal found in the Earth's crust. Major sources of the lead in air include ignition of vestige fuels and leaded gasoline, paint, smelters (metal refineries) and battery manufacturing.

VEHICLES DATASET:

Data for number of vehicles registered in different regions of Bangalore is collected for different type of vehicles namely:

12. *Number of Transport Vehicles Registered*
13. *Number of Multiaxled Vehicles Registered*
14. *Number of Lmv goods Vehicles Registered*
15. *Number of Buses Registered*
16. *Number of Taxies Registered*
17. *Number of Light motor Vehicles Registered*
18. *Total number of vehicles Registered*
19. *Number of Non transport vehicles Registered*

B. Exploratory Anaylysis:

Exploratory analysis is performed on the dataset to identify inconsistent data.It is also used to identify missing values in the given dataset.Outliners check is performed by plotting boxplot of various independent variables.

C. Data Cleaning:

Second stage in our project is data cleaning. One of the most important steps in any machine learning model is data manipulation.

Inconsistent data: Out of 20 stations 10 stations had data for less than 12 months such stations were discarded and we were left with data only for 10 stations.

Missing values*:* Since, the collected data has large number of missing values we can either delete such rows or we can populate these values with appropriate methods. In our dataset we have replaced the missing values by grouping the data based on different stations and evaluating the mean.

Forecast various independent variables using auto regressive integrated moving average model

ARIMA is an Auto Regressive Integrated Moving Average. It is a class of model that takes a collection of various standard temporal structures in time series data.

For predicting values for the next 6 months using linear regression initially we need to forecast the values of the independent variables such as pm10, pm2.5, no2, nh3, co, pb, number of transport vehicles registered, number of non-transport vehicles registered etc. on which the target parameter i.e. AQI depends.

Since the predictor’s of AQI in our dataset are independent of each other and are only dependent on the date. We use ARIMA model on the cleaned data set to forecast future 6 months data starting from Dec 2018 to May 2019 for all the pollutants and vehicle types. After populating all the predictor’s we can use linear regression model to predict the AQI for each station.

D. Linear Regression Model:

Split the data set into training and test data with a split ration of 80-20. Step wise regression is used for variable selection to identify the significant variables to be used in the model.

STEP WISE: In measurements, step wise regression is a technique of fitting regression models in which the choice of predictive parameters is carried out by a programmed procedure. In every step, a variable is measured for addition to or subtraction from the set of descriptive variables based on certain pre-quantified condition. The output of stepwise function in our development is a combination of 4 variables they are pm10, nh3, no2 and number of transport vehicles registered. On checked the combination of variables returned by step wise for P values the p values are less than 0.05 Therefore, we will use can use this model for predicting AQI. AQI is predicted for the test dataset and the accuracy of the model is analyzed using various performance estimators like root mean square error, absolute mean error etc.

E. Checking for Assumptions:

Finally we should check if the assumptions of linear regression are met.

1) Normality of Residuals: The first assumption that we need to check is the normality check of residual. Residual is the difference between the predicted and accurate value. On plotting the residual values of the training data we get a normalized bell shaped curve with mean at the origin. Therefore, the first assumption is verified.

2) Another assumption that we should check for is multi colinearity check, the independent variables should not be too correlated. When the independent variables are highly correlated we have an issue called multi colinearity. The correlation between the independent variables in our model is less than 30%.Therefore, even this assumption is satisfied.

3) Residuals should have constant variants: Another assumption that we should check for is that the residual should have constant variance, this is called homoscedasticity. When we plot the residuals against the predicted values, there is no pattern. Therefore this assumption is satisfied.

4) Check for linearity: When we plot the residual values against different independent variables we observe a random variation for each of these plots. Therefore, the model is linear.

F. Python and Jupyter Notebook:

This project is implemented using python and jupyter notebook.

1) ***Python:*** Python is a translator, top-level, general- purpose programming language. Python has a design philosophy that gives emphasis to code readability, particularly using important white space. It provides and constructs, that enable clear programming on both small and large measures.
2) ***Jupyter Notebook:*** The Jupyter Notebook is a free-source web application that permits you to create and share the documents that contain live code, comparisons, visualizations and descriptive information. Usages include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and many more.

Retrieval Number: B2437078219/19©BEIESP
DOI: 10.35940/ijrte.B2437.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IV. RESULTS

Efficiency of the given model is analyzed using 4 performance estimators they are:

ROOT MEAN SQUARE ERROR (RMSE):

Root Mean Square Error (RMSE) is a standard deviation of residuals (prediction errors). RMSE measures, how to extent these residuals. RMSE value observed for this model is **3.44.**

MIN MAX ACCURACY: is a good metric that studies the average between minimum and maximum prediction. Observed value of the min-max accuracy is **0.97**.

MEAN ABSOLUTE ERROR PERCENTAGE: The MAPE

(Mean Absolute Percent Error) calculates the size of the error in percentage terms. Mean absolute error percentage observed is **3.02%.**

MEAN ABSOLUTE ERROR: The mean absolute error (MAE) is very simple regression error metric to comprehend. We will calculate the residual for every data point, taking only the absolute value of each, so that negative and positive residuals do not cancel out. We then take the average of all these residuals. Observed value of Mean absolute error is **2.43**

V. ANALYSIS

It is observed that the top 3 stations with highest Air Quality Index are:

1) Central silk board
2) ITPL Whitefield
3) City railway station

It is also observed that the high value of AQI in Central silk board and ITPL Whitefield is a result of high value of pm10 and nh3.

Whereas, In City railway station the number of transport vehicles registered is quite high compared to other stations. Some approaches that can be used to reduce AQI in such regions are:

- Decrease diesel emissions by substituting older engines with newfangled and cleaner engines.
- Prohibition on 10 year old commercial vehicles.
- Diesel vehicles, comprising trucks, are a crucial source of fine particles. By-passing of trucks through the proposed peripheral ring road around Bangalore, etc.

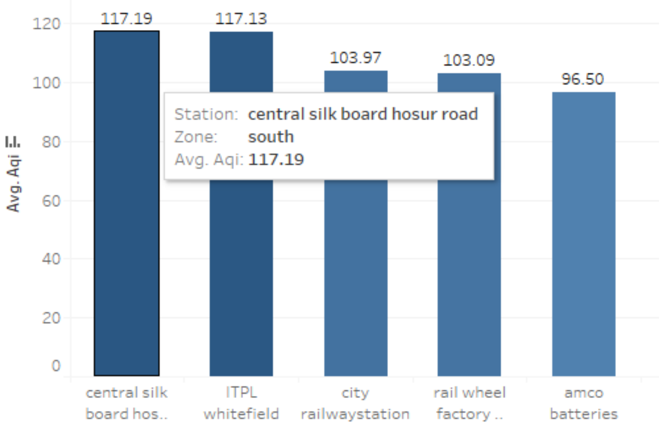
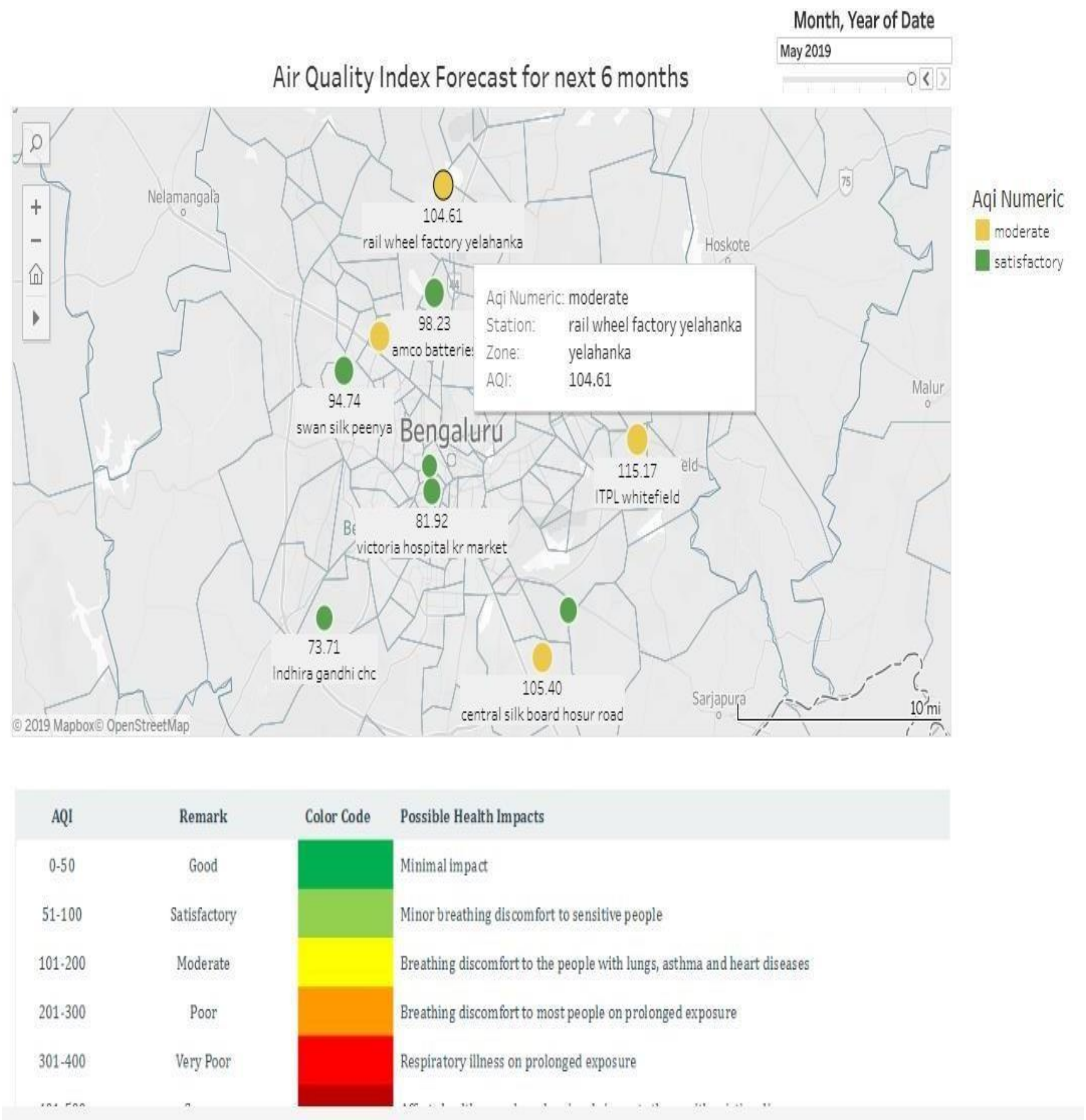


Figure 4: Top 5 stations with highest values of AQI

VI. CONCLUSION

Predicted values of the next six months of the AQI are visualized on the tableau public. Accuracy of this model is quite good. The predicted AQI has an accuracy of 96%. Future enhancements include increasing the scope of region and to include as many regions as possible currently this project aims at predicting the AQI values of different regions of Bangalore. Further, by using data of different cities the scope of this project can be expended to predict AQI for other cities as well.



Aqi Numeric: moderate
Station: rail wheel factory yelahanka
Zone: yelahanka
AQI: 104.61

AQI	Remark	Color Code	Possible Health Impacts
0-50	Good	🟩	Minimal impact
51-100	Satisfactory	🟩	Minor breathing discomfort to sensitive people
101-200	Moderate	🟨	Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor	🟧	Breathing discomfort to most people on prolonged exposure
301-400	Very Poor	🟥	Respiratory illness on prolonged exposure

Figure 5: Published dashboard on tableau public with predicted AQI value

REFERENCES

1. Qian Di, Petros Koutrakis and Joel Schwartz. “A hybrid prediction model for PM2.5 mass and components using a chemical transport model and land use regression.” In Atmospheric Environment, Volume 131, pp. 390399, 2016.
2. Ni, X.Y.; Huang, H.; Du, W.P. “Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data.” Atmos. Environ. 2017, 150, 146–161.
3. G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," Environ. Model. Softw., vol. 80, pp. 259-264, 2016.
4. PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG,” Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations,” IEEE ACCESS July 30, 2018.Digital Object Identifier 10.1109/ACCESS.2018.2849820.
5. Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie,” Air Quality Prediction: Big Data and Machine Learning Approaches,” International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018.
6. Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner and Daniel Price, “Air Quality Simulations using Big Data Programming Models,” IEEE Second International Conference on Big Data Computing Service and Applications,2016.

AUTHORS PROFILE

Mrs. Ambika G.N , BE, M.Tech (PhD) working as an Assistant Professor in the Dept of CSE, BMS Institute of Technology , Bangalore. Published 12 research papers in various areas. Currently Working in Artificial intelligence and Machine Learning

Mr. Bhanu Pratap Singh, BE, CSE, BMSIT & M.

Ms. Bhavya Sah, BE, CSE, BMSIT & M

Ms. Dishi Tiwari, BE, CSE, BMSIT & M