Detection and Prediction of Air Pollution using Machine Learning Models

Aditya C R[#1], Chandana R Deshmukh[*2], Nayana D K[*3], Praveen Gandhi Vidyavastu[*4]
[#]*Associate Professor, Department of Computer Science and Engineering,*
[*]*B.E. Student,Department of Computer Science and Engineering,*
Vidya Vikas Institute of Engineering and Technology,Mysuru, Karnataka, India 570028

Abstract—*In the populated and developing countries, governments consider the regulation of air as a major task. The meteorological and traffic factors, burning of fossil fuels, industrial parameters such as power plant emissions play significant roles in air pollution. Among all the particulate matter that determine the quality of the air, Particulate matter (PM 2.5) needs more attention. When it's level is high in the air, it causes serious issues on people's health. Hence, controlling it by constantly keeping a check on its level in the air is important. In this paper, Logistic regression is employed to detect whether a data sample is either polluted or not polluted. Autoregression is employed to predict future values of PM2.5 based on the previous PM2.5 readings. Knowledge of level of PM2.5 in nearing years, month or week, enables us to reduce its level to lesser than the harmful range. This system attempts to predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city.*

Keywords — *Pollution detection, Pollution Prediction, Logistic Regression, Linear Regression, Autoregression*

I. INTRODUCTION

Particulate matter can be either human-made or naturally occur.Some examples include dust, ash and sea-spray. Particulate matter (including soot) is emitted during the combustion of solid and liquid fuels, such as for power generation, domestic heating and in vehicle engines. Particulate matter varies in size (i.e. the diameter or width of the particle). $PM_{2.5}$ refers to the mass per cubic meter of air of particles with a size (diameter) generally less than 2.5 micrometers (μm)[13]. $PM_{2.5}$ is also known as fine particulate matter (2.5 micrometers is one 400^{th} of a millimeter). Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high [1]. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. Different machine learning models have been applied to detect air pollution and predict PM2.5 levels based on a data set consisting of daily atmospheric conditions. Naive Bayes classification and support vector machine algorithms were applied by Dan Wei [2] to get the minimum error with respect to prediction of the air quality in Beijing city. A fuzzy inference system was introduced by José Juan Carbajal et al.[4] to perform parameter classification using a reasoning process and integrating them into an air quality index.

Due to the uncertainty of the specific number PM2.5 level, the problem is simplified to be a binary classification one, that is to classify the PM2.5 level into "High" (> 115 ug/m3) and "low" (<= 115 ug/m3). The value is chosen based on the Air Quality Level standard, which set 115 ug/m3 to be mild level pollution [2].

The existing systems [2, 4, 14, 15] detect the air quality of a particular city selected by the user and groups it into different categories like good, satisfactory, moderate, poor, very poor, severe based on AQI (Air Quality Index). The data is displayed on a monthly, weekly or daily basis. Also, once the values are forecasted, the values do not change with respect to the sudden change in the atmospheric conditions or unexpected increase in traffic. The values are detected for the whole city, and cannot be verified for the accuracy of the forecasted values afterward.

There are applications that display the real-time PM2.5 levels, while some show the forecast of a particular day. However, PM2.5 levels for dates after a week is not forecasted.

This system exploits machine learning models to detect and predict PM2.5 levels based on a data set consisting of atmospheric conditions in a specific city.

The proposed system does two tasks (i). Detects the levels of PM2.5 based on given atmospheric values.

(ii) Predicts the level of PM2.5 for a particular date. Logistic regression is employed to detect whether a data sample is either polluted or not polluted. Autoregression is employed to predict future values of PM2.5 based on the previous PM2.5 readings. The primary goal is to predict air pollution level in City with the ground data set.

II. METHODOLOGY

There are two primary phases in the system:

1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly.
2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked.

And therefore, the data that is used to train the model or test it, has to be appropriate.The system is designed to detect and predict PM2.5 level and hence appropriate algorithms must be used to do the two different tasks.

Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

DATA SET

The data set that is used to train the system to detect the air quality was obtained from UCI repository [16]. The data set was to have the following attributes:

1. Temperature
2. Wind speed
3. Dewpoint
4. Pressure
5. PM2.5 Concentration(ug/m^3)
6. Result – data sample is classified either to be polluted or not polluted.

The followingplot shows that all the featuresthat are considered for the prediction are correlated and thus can be considered to train the model.

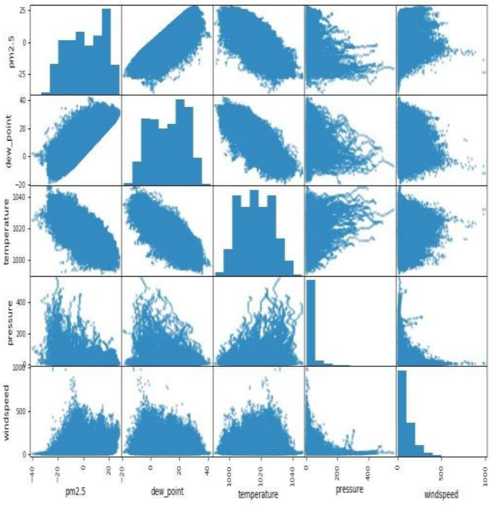

Fig.5.1 Scatter plot for the relation among attributes

A. PM2.5 level detection using logistic regression

Logistic regression [17] is the algorithm employed to detect a user-defined sample to be polluted or not. Logistic regression is the appropriate regression model to conduct analysis when the dependent variable is dichotomous (binary or has two classes). For example, here, the data set gets classified into two classes — I.E, "Polluted" or "Not Polluted". Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to explain the relationship between one or more independent variables and one dependent binary variable.

$$\textbf{Logit(p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$

Logit function is used to generate log odds of an attribute that signifies the probability of the attribute. Log odds are an alternate way of expressing probabilities, which simplifies the process of updating them with new evidence.

Based on logit function, the system classifies the training data to be either 0 (not polluted) or 1 (polluted) and verifies its accuracy using the test data. The result of the user input is also 0/1 and not the PM2.5 level.

B. PM2.5 level prediction using Autoregression

An autoregressive (AR) model considers observations from previous time steps as input to predict the value at the next time step. It is used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them. The past data is used to model the behavior, hence the name autoregressive. Basically, the process is a linear regression of the data in the series against one or more past values in the same series.

In an AR model, the value of the outcome variable (Y) at some point in time 't' is — like "regular" linear regression — directly related to the predictor variable (X). Where simple linear regression and AR models differ, is when Y is dependent on X and previous values for Y.

The AR (p) model is defined by the equation:
$y_{tk} = \delta + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-1} + A_t$
Where:

$y_{t-1}, y_{t-2} \ldots y_{t-p}$ are the past series values (lags),
A_t is white noise (i.e. randomness),
And δ is defined by the following equation:

$$\delta = \left(1 - \sum_{i=1}^{p} \phi_i\right) \mu ,$$

Where μ is the process mean

To enable auto-regression, the data set was modified into time series data set that was derived by taking the date and previous PM2.5 from the main data set. The data set used has two fields -

1. Timestamp - date and time for the data sample collected
2. PM2.5 - PM2.5 concentration (ug/m^3)

III. RESULTS

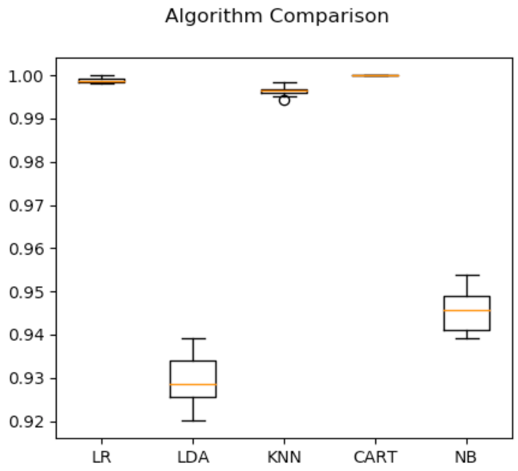

Fig.3.1 Algorithm Comparison

(i)As shown in the fig.3.1, when compared to other machine learning models applied on the data set, Logistic Regression suits the best for this system with the mean accuracy and standard deviation accuracy to be 0.998859 and 0.000612 respectively. Hence, Logistic regression can be used to clearly classify and distinguish the PM2.5 value generated based on the given sample atmospheric conditions to be polluted or not.

(ii)Autoregression applied on time series data set to predict the PM2.5 value 7 days prior to the current date, produced the Mean Squared Error(MSE) to be 27.00. MSE can be reduced by decreasing the difference between the current date and the date on which the value of PM2.5 is to be predicted.

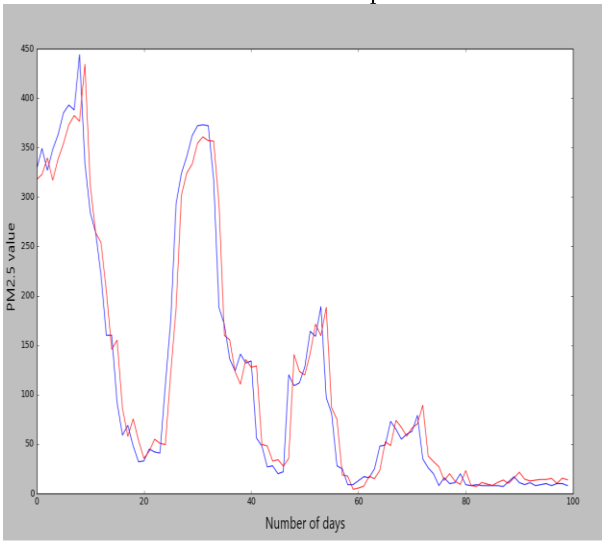

Fig.3.2 The graph depicts the actual and predicted PM2.5 value when Autoregression model was applied on the data set

As in the above fig.3.2, the graph indicates the actual values in blue color and the predicted values by red. The graph was plotted for 100 values that were split into test data.

IV. CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it.

The results show that machine learning models (logistic regression and autoregression) can be efficiently used to detect the quality of air and predict the level of PM2.5 in the future.

The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities.

REFERENCES

[1] Pandey, Gaurav, Bin Zhang, and Le Jian. &quot; Predicting sub-micron air pollution indicators: a machine learning approach.&quot ; Environmental Science: Processes & amp; Impacts 15.5 (2013): 996-1005.

[2] Dan wei: Predicting air pollution level in a specific city [2014]

[3] Dixian Zhu, Changjie Cai, Tianbao Yang and Xun Zhou: A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big data and cognitive computing [2018].

[4] José Juan Carbajal-Hernándezab Luis P.Sánchez-Fernándeza Jesús A.Carrasco-OchoabJosé Fco.Martínez-Trinidadb: Assessment and prediction of air quality using fuzzy logic and autoregressive models: Center of Computer Research – National Polytechnic Institute, Av. Juan de Dios Bátiz S/N, Gustavo A. Madero, Col. Nueva. Industrial Vallejo, 07738 México D.F., Mexico1. (2012) Doi :https://doi.org/10.1016/j.atmosenv.2012.06.004

[5] Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai : An Empirical Study of PM2.5 Forecasting Using neural network. IEEE Smart World Congress, At San Francisco, USA [2017]

[6] Athanasiadis, Ioannis N., et al. &quot;Applying machine learning techniques on air quality data for real-time decision support.&quot; First international NAISO symposium on information technologies in environmental engineering (ITEE&#39;2003), Gdansk, Poland. 2003.

[7] Ioannis N. Athanasiadis, Kostas D. Karatzas and Pericles A. Mitkas. &quot;Classification techniques for air quality forecasting.&quot; Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.

[8] M. Caselli &amp; L. Trizio &amp; G. de Gennaro &amp; P. Ielpo. &quot;A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model.&quot; Water Air Soil Pollut (2009) 201:365–377.

[9] S.Bordignon, C. Gaetan and F. Lisi, &quot;Nonlinear models for ground-level ozone forecasting.&quot; Statistical Methods and Applications, 11, 227-246, (2002).

[10] K.Chidananda Gowda and Edwin Diday. Symbolic clustering using a new dissimilarity measure. pattern recognition, 24(6):567–578, 1991.

[11] K.Chidananda Gowda and Edwin Diday. Symbolic clustering using a new similarity measure. IEEE Transactions on Systems, Man, and Cybernetics, 22(2):368–378, 1992.

[12] Edwin Diday. Symbolic data analysis: a mathematical framework and tool for data mining. In Advances in Data Science and Classification, pages 409–416. Springer, 1998.

[13] https://en.wikipedia.org/wiki/Particulates

[14] http://aqicn.org/city/india

[15] https://app.cpcbccr.com/AQI_India/

[16] https://archive.ics.uci.edu/ml/data sets/Air+quality

[17] Source code for logistic regression: https://github.com/scikit-learn/scikit-learn/blob/a24c8b46/sklearn/linear_model/logistic.py#L962