International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 4, ISSN No. 2455-2143, Pages 420-426
Published Online August 2020 in IJEAST (http://www.ijeast.com)

FORECASTING AIR QUALITY USING MACHINE LEARNING

Shumile Khan, Manoj Yadav
Department of Computer science & engineering
Al Falah University
Faridabad, Haryana, India

Abstract— **Prediction of air pollution is an increasingly important problem. There is a measurement benchmark to know the level of different pollutant parameters. Recent studies show different methods in the field of computer science and engineering to create the relationship between the concentrations of Air pollutants and their emission sources. We will develop artificial neural network model to train the system based on some data and will use auto regression method for further processing. We will implement a system for predicting the data that is trained efficiently. For this, we will collect air pollution data (pollutant parameters) and based on this data we will train the system to predict the air pollutants accurately. Our system would be fit for forecasting air quality level of any city or area based on previously trained data of air pollutants.**

Keywords— **Dataset, ANN, Auto Regression, K-Means clustering, Adam Optimizer, Confusion Matrix**

I. INTRODUCTION

Air Quality has been a reason to worry for many decades with the rapidly increasing pollutant concentration in developing cities. There is a standard measurement of Air Quality; based on these quality levels, air pollutants are recorded. These air quality levels show the pollutant concentrations and in respect to their scales, features and severity they impose major health issues. Air Quality Index (AQI) is the benchmark of air quality that estimates air conditions based on concentration of air pollutants. Most of the air pollutants include sulphur dioxide (SO2), ozone (O3), carbon monoxide (CO), nitrogen oxides (NOx), particulate matter (PM), pesticides, and many more[1]. Due to increased air pollutants concentrations, Sick and death rates have been multiplying in association with it. It is very difficult to find any pollutant in the human respiratory system as if we talk about PM 2.5; it is a tiny pollutant particle which is about 2.5 µm diameters and it is very difficult to expel pm 2.5 from our body[2]. Over the past decades, there has been a vast change in population. In result; number of urbanization and industrialization are increased in many cities that led to more air pollution that caused a negative impact on indoor and outdoor air quality [3].

As we know troposphere is the lowest layer of Earth's atmosphere. This layer has the air we breathe and it contains 78.08% nitrogen, 20.95% oxygen, 0.93% argon, 0.04% carbon dioxide, and small amounts of other gases. Air also contains a small amount of water vapor. The depth of lower layer of troposphere often depends on solar radiation and temperature. So through convection process temperature shows a less influence on air pollutants. It increased an air pollution level that depends on direction of wind. Weather humidity is also a cause for making fine particulates heavier. So it is unrealizable to depend on one parameter to understand air pollution because topographical characteristics are not the same and also not same kind of people or things [4].

Most of the researchers worked on forecasting air quality or air pollution based on air pollutants, methods, connection between air pollution and human health. But we found that there are fewer researches which focused to control air pollution and yet there is no successful way to control air pollution. We can protect us by wearing a mask on our mouth but it is not a healthy way because of the excessive carbon dioxide intake.

In previous research it is shown that the machine learning technology has been impactful for forecasting air quality. That’s why it has been mostly using by many researchers. Machine learning is a very efficient method to forecast air pollution, so we propose an idea for applying machine learning approach to forecast air quality levels. However, the effect of temperature, pressure, and even the amount of solar radiation does not vary throughout the time period and wind direction and speed depend on the topographic features [4]. These are the biggest challenges in predicting weather and air quality.

In this work, we will train our system for efficiently forecasting air quality level based on some previous data of air quality. We are focusing on the improved performance of generalization behavior of ANN model and to train a complex model with advanced optimization algorithm.

II. RELATED WORK

Previous studies have shown their work methodology by applying machine learning approach to forecast air quality levels. We will go through some work.

Kalapanidas et al. (2001) made effects on air pollution based on meteorological parameters (solar radiation, humidity, temperature etc.) by using case-based reasoning (CBR) system and generalized air pollution into distinct levels such as low, med, high, alarm [5].

Athanasiadis et al. (2007) used s-fuzzy lattice neuro computing classifier. They forecasted ozone concentration and classified into three levels. Their prediction was based on some gases and meteorological parameters [6].

Yu Zheng et al. (2013) worked on urban air pollution. They Predicted the concentration of tiny particle PM 2.5 on the basis of other factors as well such as land uses, meteorological variables etc. There was limited air monitoring station and they faced the issue of non-linear urban spaces [7].

Yu Zheng et al. (2015) implemented a semi-supervised inference model in terms of many other factors for instance human mobility, city dynamics, meteorology etc. They also proposed entropy minimization model to show the appropriate location. Advantage of this model are to minimize the inference error, minimized the model uncertainty [8].

Zheng et al. (2015) used a data driven approach that involves current meteorological parameters, temperature and so on. They took 48 hours air quality data reading and made their predictions. There are four major components of their prediction work:
1) They used a linear regression-based temporal model to analyze the local factors of air quality.
2) They made a spatial predictor based on neural network to analyze global factors.
3) They used a dynamic aggregator that combined the predictions of the spatial and temporal model.
4) They made an inflection predictor to capture sudden changes in air quality [9].

Li et al. (2011) in their research they found the concentration of particulate matter PM 2.5 by using spatio-temporal interpolation methods. Their assessment was based on distinct accuracies of interpolation results and then they selected the most effective one to perform PM 2.5 interpolations [10].

Masood et al. (2020) in their work they developed SVM and ANN machine learning model and compared them. They analyzed the performance of both models based on data parameters and other factors and provided the result that the ANN model gives the better accuracy than SVM model [11].

Lin et al. (2018) in their work they selected the air pollutant concentration data from 2005 to 2014. After some period of time, new indicators were added in record. Their research was focused on status of industrial air pollution control and handled slightly the control status of other pollution sources [12].

III. Data Observation

A. Data Collection -

In this work, research methodology works on data and Machine learning models. Various air pollutants data concentration such as NO (Nitric Oxide), NO2 (Nitrogen dioxide), NOx (Oxides of Nitrogen), CO (Carbon Monoxide), SO2 (Sulphur dioxide), NH3 (Ammonia) etc. and many meteorological data like temperature, wind speed, wind direction, humidity etc. have been utilized to develop the model for predicting air pollution [13]. We need to do ensure that data should be in required manner to train learning model accurately. Usually if we have huge data, the challenges will be more difficult.

We took essential data from the online source; this raw data contains many error and null values. We processed it through many steps that are necessarily taken to process data. The null values, inconsistencies should be excluded from the data.

B. Preprocessing-

Data preprocessing is a machine learning technique that improves the quality of data because the raw data we are collecting is not suitable for further analysis. So data preprocessing helps to clean and organize the raw data. Hence this raw data is going to be clean through various steps. These steps are: data selection, data preprocessing and data transformation [14].

C. Clustering-

Once we get the dataset we need to check whether the dataset have labeled are not. Since our data set don't have label so we need to cluster them because then only we can give the label. For that we are using K-Means clustering algorithm for grouping the data points with five clusters. Once we get labeled dataset we need to train our classification algorithm.

1. K-Means clustering:

It is an unsupervised machine learning algorithm. It is used to make groups of similar data points (clusters) from unlabeled data; k is the fixed number of clusters. In this algorithm, centroid is a data point at the center of the cluster; it may be real and imaginary [15]. In a cluster, data points may be more so we need to find the mean value by using this algorithm means different data points will be assigned to a cluster and finds the distance between the centroid and data points. And the resultant distance value (mean value) should be minimal. We are using scikit-learn package for implementing k-means clustering.

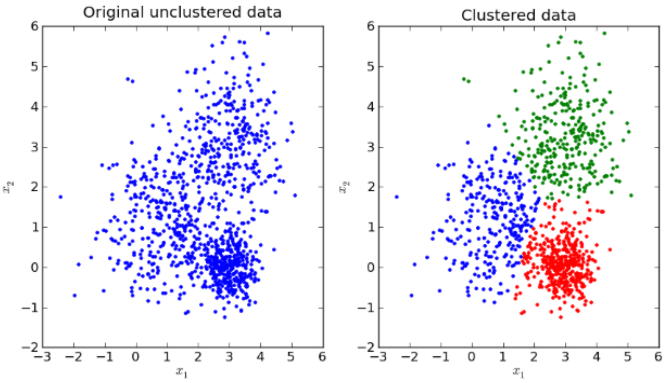


fig.1. Unclustered and Clustered data

As a quick review from this module, we took data from the online source. Keep data in csv format. Then we analyze the dataset contain any null values are not, apply scikit-learn library, we will do normalize data with it and make it ready for ANN algorithm.

IV. Machine Learning Methodology

There are two primary phases in this methodology. First is Training phase. In this phase, we are constructing a model based on artificial neural network and trained by using various features and a dataset. Second is Testing phase. In this, the model is provided with the data inputs and is tested for future work. The data that is used to train or test the model has to be appropriate. The system is designed to detect and predict air quality level; 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 and also meteorological variables which display explicit characteristics on different time variants. Pollutant concentrations are easily influenced by these Parameters and always get dense in nature. In an article, it is mentioned that "the highest air pollutant densities in Istanbul are measured not only in summer months due to high temperatures and evaporation but also in winter months due to high level of gasoline consumption" [16]. So different parameters may lead to air pollutants get denser.

Hence it will create many problems if we use fully connected artificial neural network in prediction of air quality without included time variants feature. So, to rectify this we have sequential deep learning method available. In previous studies, they used basic approaches for forecasting and took time as a factor and did not used past values for prediction. Hence fundamental unit of ANN doesn't forecast with sequential methods as their connections are limited [15]. In this work, we are using Artificial Neural Network and Auto Regression model for evaluation based on various key points. The methodology outlines the proposed idea and methods that how the project works is conducted.

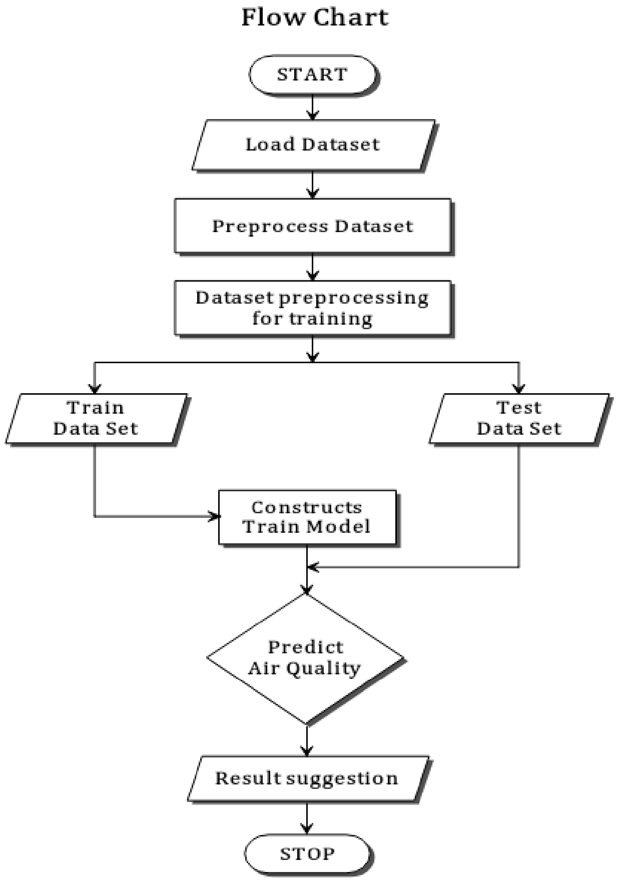


Fig.2. Flow Chart of Proposed Method

A. Model construction-

We split dataset into two parts—Train data and Test data. Then we maintained the test dataset separately and chose arbitrary x% of dataset (70%) to form the effective train data and the rest data (100-x=30%) to form the test dataset [17]. Now our dataset is ready for classification so we need to train our model with Artificial Neural Network algorithm using training data and then test accuracy with test data. The ANN model is trained with a number of accurate iteration and test this model on different test datasets. If we satisfy with the accuracy of ANN model, we can forecast the new record which is generated by Auto Regression.

B. Artificial Neural Network-

An artificial neural network (ANN) is a special architecture of feed forward neural network. In this step, we will proceed with calculating all values for hidden layers and output layers in ANN.

- Assign data values to input nodes.
- Then we will calculate all hidden values.

$H_1 = I_1W_1 + I_2W_3 + B_1W_5$

$H_2 = I_1W_2 + I_2W_4 + B_1W_6$

- Take sigmoid function for hidden layer for activation.
  $S(x) = 1/1+e^{-x}$
- Calculate activation values for hidden nodes.
  $HA_1 = 1/1+e^{-H}{}_1$
  $HA_2 = 1/1+e^{-H}{}_2$
- Now we will calculate data values of output nodes.
  $O_1 = HA_1W_7 + HA_2W_9 + B_2W_{11}$
  $O_2 = HA_1W_8 + HA_2W_{10} + B_2W_{12}$
- Take linear function (activation function) for the output layer.
  $Y = f(x) = x$
- Now calculate activation values for output nodes.
  $OA_1 = O_1$
  $OA_2 = O_2$
- In the end we will calculate total error. In the following equation y_i is the desired output and OA_i is the obtained value for node i.
  $e = \frac{1}{n}\sum_{i=1}^{2}(yi - OAi)^2$ [18]

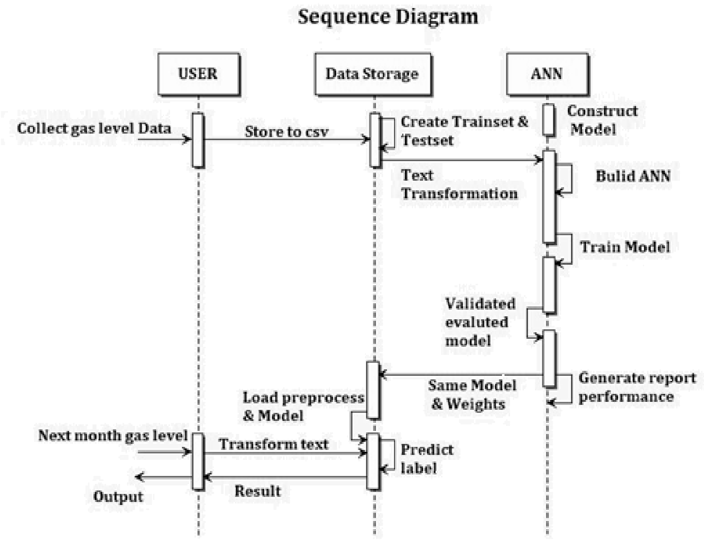


Fig. 4. Sequance Model Diagram

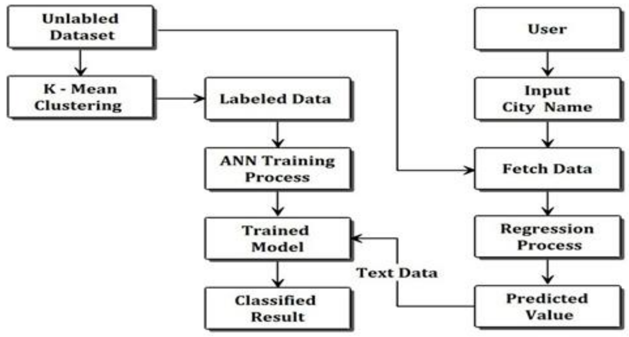


Fig.3. System Architecture

C. Sequence Model-

It helps to arrange all layers (from input to output) in a stack format. Sequential model (method) doesn't provide a feature of having multiple input and output or sharing the layers with one another in ANN model [19]. We can say simply it is a kind of unusual model that allow us to arrange all layers in a sequential manner. We just need to draw on a sequential model and define all the layers from input to output. To construct this; we import a sequential model package from keras.

D. Model Evaluation-

Evaluation is an essential feature in the process of model development. It is used to choose an accurate model that can easily represent data in an organized manner and also for further use. Evaluation of a model performance can be a critical task because it involves data that we usually take for training. When any model works well on training dataset and unknown data too; then the chances of over-fitting may be increase if this learning model starts to memorize data instead of learning. To prevent the occurrence of over-fitting, hold out and cross validation methods are used for evaluating model performance [20].

Following Steps are included:

- Analyze and create Training dataset.
- Analyze the need for test data.
- At the time of test data design phase, we need to analyze test data thoroughly.
- Create test data.
- Execute tests.
- Save data.

1. Use of evaluate function:

It is a built-in function that is used for evaluation of a specified python expression. It passes the expression as a string and gives the output as an integer.

For example: eval ("1 + 1") interprets and executes the expression "1 + 1" and returns the result "2".

E. Auto Regression-

It is a modeling algorithm that is used for predicting time series data. It uses previous observations into a regression equation as input for forecasting new data. We express it as AR (x), where x shows the number of lagged values.

Y = c0 + c1*X1

By using above equation, we can substitute the values of coefficients in c0, c1 variable (the values we get from trained data by optimization), Y is the prediction and X1 is the input variable.

We can understand this equation by following example of Auto Regression model:

Y= c0 + c1*X (t-1) + c2*X (t-2) + c3*X (t-3)

Where X (t-n) is the input variable at previous time series of the dataset [21].

V. Result

A. Confusion Matrix-

It is a machine learning method that allows us to predicting the result in an accurate and precise manner. Confusion matrix is a table to display the result of a classification model. It is used when we have two or more classes data output and we need to find out what are the actual values and predicted values. It is just a confusion avoidance solution when classification model gets confused while predicting the result.

Table -1 Confusion Matrix Table

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Table 1 shows class 1 and class 2 is positive and negative respectively. Now in order to understand above classes and observations; we need to know what are TP, FP, TN and FN.

TP is True Positive in terms of observation is predicted positive and it is true.

FP is False Positive in terms of observation is predicted positive and it is false.

FN is False Negative in terms of observation is predicted negative and it is false.

TN is True Negative in terms of observation is predicted negative and it is true [22].

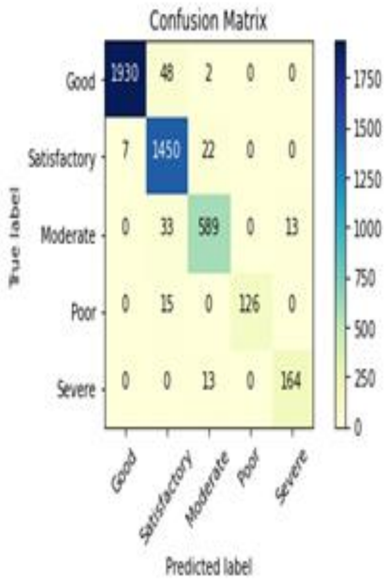
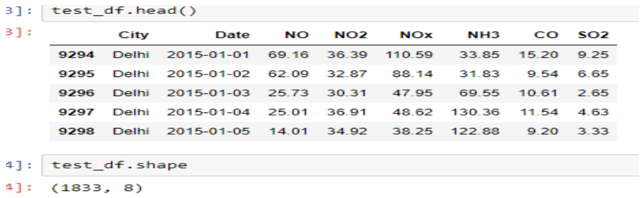


Fig.5. Test data Records

As shown in the fig5, we plotted the predicted result (Test data) in confusion matrix. We labeled the test data as Good, Satisfactory, Moderate, Poor, Severe Air quality based on testing phase. This matrix shows the number of data records i.e. 1930 records that are good, 1450 data records are satisfactory and 589 are Moderate and rest we can see in matrix. We took around 70 % of data in training and rest we put for testing to analyze our model is appropriately working or not.

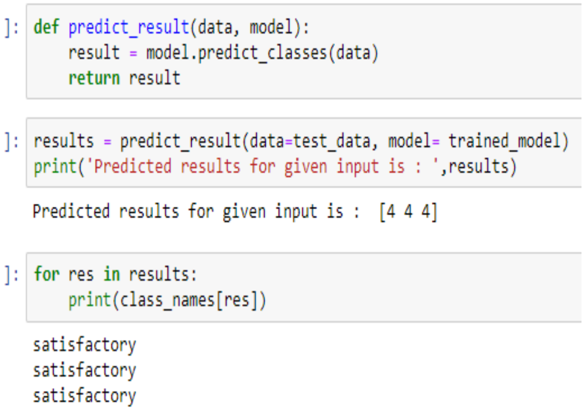


```
3]: test_df.head()
3]:
           City        Date     NO    NO2     NOx     NH3     CO   SO2
     9294  Delhi  2015-01-01  69.16  36.39  110.59   33.85  15.20  9.25
     9295  Delhi  2015-01-02  62.09  32.87   88.14   31.83   9.54  6.65
     9296  Delhi  2015-01-03  25.73  30.31   47.95   69.55  10.61  2.65
     9297  Delhi  2015-01-04  25.01  36.91   48.62  130.36  11.54  4.63
     9298  Delhi  2015-01-05  14.01  34.92   38.25  122.88   9.20  3.33

4]: test_df.shape
4]: (1833, 8)
```

Fig.6. Output of test data

Then we took a new dataset for predicting the future air quality of certain time (next three month). And according to our trained ANN model, we found the forecasted result as satisfactory air quality.



```
]: def predict_result(data, model):
       result = model.predict_classes(data)
       return result

]: results = predict_result(data=test_data, model= trained_model)
   print('Predicted results for given input is : ',results)

   Predicted results for given input is :  [4 4 4]

]: for res in results:
       print(class_names[res])

   satisfactory
   satisfactory
   satisfactory
```

Fig.7. Predicted Result of next three month

VI. CONCLUSION AND FUTURE WORK

We have created a capable machine learning model that can forecast air quality. We are using ANN and regression learning to predict the air pollutants accurately. With this model, we can forecast the Air Quality and take the proper precaution to stay away from poor air quality and also we can implement various effective ways to fight with this. It is a progressive learning model for future perspective. This model will analyze all the pollution parameters and build a prediction model involving all the other factors which exhibit better performance**.**

In this work, the efficiency of model was analyzed with different factors. According to confusion metrics our study reached to different labeled classes for positive and negative observations. Under same parameters, artificial neural network and auto regression model performed efficiently while predicting air pollution.

ANN model is efficiently trained so it can forecast air quality of different cities or areas. And also it can be used in other applications as well. We also have some objectives for future work. We analyzed most of the population spend their more time in indoor environments. The indoor environments of the study area are associated with higher concentration of PM2.5 and PM10 and these small pollutants can increase other diseases that may cause health problems severely.

We will generalize the dust measurement which would be sufficient to relate the PM2.5 concentration and possible health effects involving other elements concentration. Our specific objective would be to collect data on the basis of different time series that changes hourly and daily, meteorological parameters and prediction of PM2.5 and PM10 specifically, these are tiny particles that causes a major impact in human respiratory system and maintaining good health and eyesight. We will implement more enhanced model which can forecast air quality based on all necessary present parameters by using effective machine learning methods. Also we will find some effective ways to control air pollution and the cost estimation. Air pollutant concentrations rapidly increasing and became one of the major tasks as it directly affects the human health. It is important that people know what is the air quality level in their surroundings and takes a step towards fighting with it.

VII. REFERENCE

[1] Curtis, L.; Rea, W.; Smith-Willis, P.; Fenyves, E.; Pan, Y. Adverse health effects of outdoor air pollutants. Environ. Int. 2006, 32, 815–830.

[2] Xing YF, Xu YH, Shi MH, Lian YX. The impact of PM2.5 on the human respiratory system. J Thorac Dis. 2016; 8(1):E69-E74.

[3] D. Massey, J Masih, A Kulshreshtha, M Habil, A Taneja. Indoor/outdoor relationship of fine particles less than 2.5 μm (PM2.5) in residential homes locations in central Indian region, 2009, 44, 2037-2045.

[4] Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez, Yves Rybarczyk, "Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters", Journal of Electrical and Computer Engineering, vol. 2017, 14 pages, 2017.

[5] Kalapanidas, Elias & Avouris, Nikolaos. (2001). Short-term air quality prediction using a case-based classifier. Environmental Modelling & Software. 16. 263-272. 10.1016/S1364-8152(00)00072-4.

[6] Kaburlasos, Vassilis & Athanasiadis, Ioannis & Mitkas, Pericles. (2007). Fuzzy lattice reasoning (FLR) classifier and its application for ambient ozone estimation. International Journal of Approximate Reasoning. 45. 152-188. 10.1016/j.ijar.2006.08.001.

[7] Yu Zheng, Furui Liu, Hsun-Ping Hsieh, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, U-Air: when urban air quality inference meets big data, 2013, Pages 1436–1444.

[8] Hsun-Ping Hsieh, Shou-De Lin, Yu Zheng, Inferring Air Quality for Station Location Recommendation Based on Urban Big Data, 2015, Pages 437–446.

[9] Zheng, Yu & Yi, Xiuwen & Li, Ming & Li, Ruiyuan & Shan, Zhangqing & Chang, Eric & Li, Tianrui. (2015). Forecasting Fine-Grained Air Quality Based on Big Data. 2267-2276. 10.1145/2783258.2788573.

[10] Li, L. & Zhang, Xingyou & Holt, James & Tian, J. & Piltner, Reinhard. 2011, spatiotemporal interpolation methods for air pollution exposure. SARA 2011 - Proceedings of the 9th Symposium on Abstraction, Reformulation, and Approximation. 75-81.

[11] Masood, Adil & Ahmad, Kafeel. (2020). A model for particulate matter (PM2.5) prediction for Delhi based on

machine learning approaches. Procedia Computer Science. 167. 2101-2110. 10.1016/j.procs.2020.03.258.

[12] Peng Su & Degen Lin & Chen Qian, 2018. "Study on Air Pollution and Control Investment from the Perspective of the Environmental Theory Model: A Case Study in China, 2005–2014," Sustainability, MDPI, Open Access Journal, vol. 10(7), pages 1-16, June.

[13] Zhu, Dixian & Cai, Changjie & Yang, Tianbao & Zhou, Xun. (2018). a Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big Data and Cognitive Computing. 2. 5. 10.3390/bdcc2010005.

[14] Medium.com.Xenonstack/data-preparation-process-preprocessing-and-data-wrangling

[15] datacamp.com/community/tutorials/k-means clustering-Python

[16] Kaya, K., Gündüz Öğüdücü, Ş. Deep Flexible Sequential (DFS) Model for Air Pollution Forecasting. Sci Rep 10, 3346 (2020)

[17] Towardsdatascience.com/train-validation-and-test-sets

[18] Kdnuggets.com.2019/11/build-artificial-neural-network scratch

[19] Keras.io/guides. Sequential model

[20] Saedsayad.com. Model evaluation

[21] Machinelearningmastery.com.Autoregressionmodels Time-series-forecasting-python

[22] Machinelearningmastery.com/Confusion-matrix machine learning

[23] Book: Principles and practices of Air Pollution Control and Analysis by J. R. Mudakavi