IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

Prophecy of Air Quality using KNN-LSTM

Mr K. S. Raghu Kumar[1], Hemanth S[2], Swetha V[3], Sunil Naik V. S[4]
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4]
Rao Bahadur Y Mahabaleswarappa Engineering College Bellary, Karnataka, India

Abstract: *Since the vast majority of the current air quality Index i.e (AQI) foreseeing models zeroed in on forecast of the time series information of a solitary objective observing station, they neglected to consider the connection and shared impact among the air quality checking station destinations and the spatio-transient attributes of air quality. This will prompt a specific one-sidedness during air quality expectation of a specific site. A prototype to predict the AQI for a short period of time was based on the K-nearest algorithm and long short memory was proposed. The air quality index of stations was made into data sets and fed for testing processing of data in the LSTM model whose prediction accuracy was dependent on the time correlation. Long Short-term Memory Neural Network: The Recurrent Neural Network problem which involved disappearing gradient when dealing with long term dependency came to an end with development of Long Short-Term Memory (LSTM) model .The working of this special type of RNN involves adding of additional layer of memory units such that the time series controllable and maintainable by using the 3 controllers to control the more than one memory cells in the memory units. KNN algorithm: When it comes to classification in the machine learning the K nearest neighbour algorithm stands in prominence, the algorithm works by predicting the relationship and distance between the data sets or samples given of different sort..*

Keywords: LSTM, RNN, KNN

I. INTRODUCTION

In recent years, there have been more serious air pollution problems in many areas of China especially in winter, in which haze frequently invaded in many cities. Thus, monitoring and forecasting of air pollution indicators are of great importance. Intelligent Environmental Protection required researchers to rely on modern sensing technology, network technology and other means to effectively analyze and speculate environmental issues such as air quality for the purpose of providing a basis for scientific and rational decision-making. Since 2013, the air quality index (AQI) has been adopted in China to evaluate the quality of air quality and classify the city's air quality. A larger AQI value means a more serious air pollution condition. On the foundation of environmental monitoring, the evaluation and prediction of air quality have become a topic of increasing concern to scholars at home and abroad. In this study, an integral short-term air quality prediction model based on K-nearest neighbor and Long Short-Term Memory was proposed. The AQI index was forecasted using the monitoring data of the provincial control station in Beijing as a sample, to accurately understand the trend of air quality changes.

II. EXISTING SYSTEM

Experimental outcome show that it improves prediction efficiency in comparison with basic models. Air pollution has attracted huge concentration involving the everyday lifetime of men and women. It has terrible influence on human health and everyday lifestyles throughout episodes of severe air pollution with the broaden of reasons and kinds of air pollutants, the complication of pollutant attention prediction has elevated. As a result, it's imperative to make use of ecological analyzing data to more appropriately guess city air pollution levels. Conventional prediction methods, such as numerical evaluation and machine learning, are commonly used in this kind of prediction. Nevertheless, a few drawbacks of those methods were recently recognized as given below. First, numerical prediction ways are situated on knowledge as abridged with the aid of historic information or the nature of pollutant exchange.

III. LITERATURE SURVEY

Anikender Kumar, Pramila Goyal (2011) presented the study that forecasts the daily AQI value for the city Delhi, India using previous record of AQI and meteorological parameters with the help of Principal Component Regression

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

(PCR) and Multiple Linear Regression Techniques. They perform the prediction of daily AQI of the year 2006 using previous records of the year 2000-2005 and different equations. After that this predicted value then compared with observed value of AQI of 2006 for the seasons summer, Monsoon, Post Monsoon and winter using Multiple Linear Regression Technique . Principal Component Analysis is used to find the collinearity among the independent variables. The Principal components were used in Multiple Linear Regression to eliminate collinearity among the predictor variables and also reduce the number of predictors . The Principal Component Regression gives the better performance for predicting the AQI in winter season than any other seasons. In this study only meteorological parameters were considered or used while forecasting the future AQI but they have not considered the ambient air pollutants that may cause the adverse health effects.

Aditya C R (et al.2018) employed the machine algorithms to detect and forecast the PM2.5 concentration level on the basis of dataset containing atmospheric conditions in a specific city. They also predicted the PM2.5 concentration level for a particular date . First of all they classify the air as polluted or not polluted by using Logistic Regression algorithm and then Auto Regression algorithm was used to predict the future value of PM2.5 depending upon previous records.

HeidarMaleki (et al.2019) predicted the hourly concentration values for the ambient air pollutants NO2, SO2, PM10, PM2.5, CO and O3 for the stations Naderi, Havashenasi, MohiteZist and Behdasht in Ahvaz, Iran which is the most polluted city in the world. They have also calculated and predicted Air Quality Index (AQI) and Air Quality Health Index (AQHI) for the four air quality monitoring stations in Ahvaz mentioned above. They used Artificial Neural Network (ANN) machine learning algorithm for the prediction of air pollutants concentration (hourly) and two air quality indices AQI and AQHI over the August 2009 to August 2010. Input to ANN algorithms involves the factors such as meteorological parameters, Air pollutants concentration, time and date

Nidhi Sharma (et al.2018) had gone through the detailed data analysis of air pollutants from 2009-2017 and also proposed the critical observation of 2016-1017 air pollutants trend in Delhi, India . They have predicted the future trends of various pollutants as Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Suspended Particulate Matter (PM), Ozone (O3), Carbon Monoxide (CO) and Benzene. By using data analytics Time series Regression forecasting they have predicted the future values of the pollutants mentioned earlier on the of previous records. According to this study results the AnandVihar and Shadipur monitoring stations of Delhi are under the study. The result shows that there is a drastic increase in PM10 concentration level, NO2 and PM2.5 are evidently increased showing the increased pollution in Delhi [14]. CO is predicted to reduce by 0.169mg/m3 , there is increase in NO2 concentration level for coming years by 16.77 µg/m3 , Ozone is predicted to increase by 6.11mg/m3 , Benzene reduce by 1.33mg/m3 and SO2 is forecasted to increase by 1.24µg/m

Ziyue Guan and Richard O. Sinnot (2018) used the various machine learning algorithms to predict the PM2.5 concentration. Data were collected from the official website of Environment Protection Agency (EPA) for the city Melbourne that contains PM2.5 air parameter and they have also collected the unofficial data from Airbeam which is the mobile device used to measure PM2.5 value . The machine Learning Algorithms Artificial Neural Network (ANN), Linear Regression (LR) and Long Short Term Memory (LSTM) recurrent neural network were used for the PM2.5 prediction but out of these algorithms LSTM gives the best performance ad predict the high PM2.5 value with reasonable Accuracy.

Mohamed Shakir and N. Rakesh (2018) have analysed the proportion of various air pollutants (NO, NO2, CO, PM10 and SO2) with respect to the time of the day and the day of the week and estimated the effect of environmental parameters as temperature, wind speed and humidity on the air pollutants mentioned above with the help of WEKA tool . The data was collected from pollution control board of Karnataka. By using ZeroR algorithm in WEKA tool the study come up with the results that shows that the concentration levels of air pollutants increase during the working days and especially during the peak hours of the day and decrease during week-ends or holidays . Using Simple K-means Clustering algorithms the study shows the relationship or dependencies between the environmental factorslike Temperature, wind speed and humidity and the air pollutants like NO, NO2, PM10, CO and SO2.

Huixiang Liu (et al.2019) have taken two different cities Beijing and Italian city for the study purpose. They have forecasted the Air Quality Index (AQI) for the city Beijing and predicting the concentration of NOxin an Italian City depending on two different publicly available datasets. The first Dataset for the period of December 2013 to August 2018 having 1738 instances is made available from the Beijing Municipal Environmental Centre which contains the fields like hourly averaged AQI and the concentrations of PM2.5, O3, SO2, PM10, and NO2 in Beijing. The second Dataset with

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

9358 instances is collected from Italian city for the period of March 2004 to February 2005. This dataset contains the attributes as Hourly averaged concentration of CO, Non methane Hydrocarbons, Benzene, NOx, NO2 . But they focused majorly on NOxprediction as it is one of the important predictor for Air Quality evaluation. They used Support Vector Regression (SVR) and Random Forest Regression (RFR) techniques for AQI and NOx concentration prediction. SVR shows better performance in prediction of AQI while RFR gives the better performance in predicting the NOx concentration.

IV. PROBLEM STATEMENT

To estimate the air pollution level using past data-set and to
predict the increasing pollution levels after a Specific Time period. Many researchers have made great contributions to the problem of air quality prediction in recent years. Various patterns and basic trends in air quality are identified by quantitative researches combining the latest technology. The main technologies and implementation methods in these achievements include the following categories.

4.1 How does the AQI work?

Think of the AQI as a yardstick that runs from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern. For example, an AQI value of 50 or below represents good air quality, while an AQI value over 300 represents hazardous air quality.

For each pollutant an AQI value of 100 generally corresponds to an ambient air concentration that equals the level of the short-term national ambient air quality standard for protection of public health. AQI values at or below 100 are generally thought of as satisfactory. When AQI values are above 100, air quality is unhealthy: at first for certain sensitive groups of people, then for everyone as AQI values get higher.

The AQI is divided into six categories. Each category corresponds to a different level of health concern. Each category also has a specific color. The color makes it easy for people to quickly determine whether air quality is reaching unhealthy levels in their communities.

AQI Basics for Ozone and Particle Pollution

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

V. METHODOLOGY

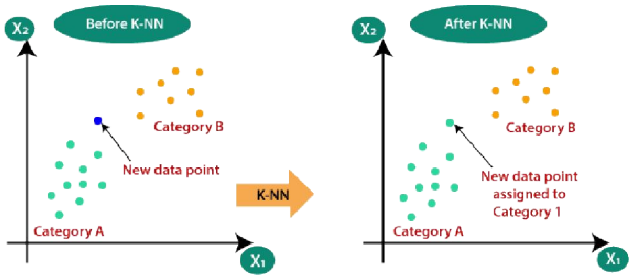
KNN Algorithm K-nearest neighbor (KNN) is one of the machine learning algorithms that can be used for classification. Its main idea was to predict the correlation between samples through calculating the distance between different types of samples. The commonly used method for measuring distance included Euclidean distance. Markov distance, Manhattan distance, etc. Since its simplicity and easiness in implementation and fast in training, our research chose Euclidean distance to calculate and measure the correlation of AQI between ness degree between the target prediction station and nearby stations and further to select the target station. Several monitoring stations with higher degrees were combined with their air quality data for prediction. The Euclidean distance is calculated with equation (1) as follows: dij = √∑(xik − xjk) 2 n k=1 （1） I represented the target monitoring station, j represented the jth monitoring station except the target monitoring station, dij represented the Euclidean distance between the two monitoring stations,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

k represented the specific time, xik and xjk represented the AQI of j monitoring station and i monitoring at k time. specific time, xik and xjk represented the AQI of j monitoring station and i monitoring at k time.

5.1 Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



5.2 How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

VI. PREDICTION BASED ON CLASSIIC MACHINE LEARNING METHODS

Classic machine learning techniques, such as regression analysis, principal component analysis, BP (Back Propagation) network and artificial neural network, were once the mainstream method of air quality prediction. Petr and Vladimir designed a model based on feed-forward neural networks of perceptron and fuzzy inference systems for air quality prediction. Kang and Qu established BP neural network based on genetic simulated annealing algorithm optimization to predict air quality. Wang et al. trained a BP neural network based on the historical monitoring data of air pollutants to predict PM2.5 mass concentration. Rajput and Sharma represented the variation of AQI (Air Quality Index) with a multivariate regression model. Major parameters, such as ambient temperature, relative humidity and bar pressure, were considered in the regression model for AQI computation. Mahajan et al. clustered the monitoring stations based on the geographical distance to reduce the forecasting errors and achieve acceptable forecast results of PM2.5 concentrations. Li et al. built a dynamic evaluation model for forecasting the air quality data based on the fuzzy mathematical synthetic evaluation. The future air quality status would be built by the fuzzy synthetic assessment model based on entropy weighing method and whose results showed that the proposed evaluation model is a practical tool. These classic methods and models have some advantages, such as simple algorithms, easy-to-understand processing and acceptable prediction results. These methods can well predict the trend of air quality changes. However, it is difficult to obtain specific accurate forecast values of air quality.

VII. PREDICTION BASED ON LSTM MODEL

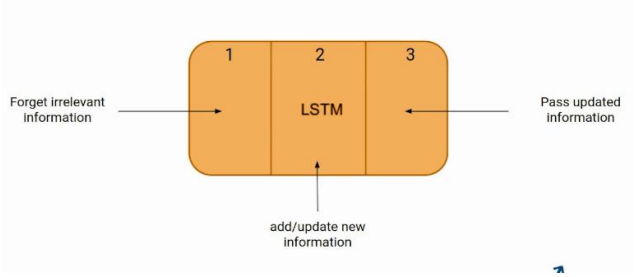
The next type of methods are the LSTM-based prediction models. LSTM is an improved algorithm of RNN network, which can memorize long-term information in sequence data. Song et al. proposed LSTM-Kalman time prediction model. The model stores the information contained in the pre-order data by using LSTM, while adjusts the basic time data sequence by Kalman filtering. Wang et al. established CT-LSTM by combining CT (Chi-square Test) and LSTM. CT is used to determine the influencing factors of air quality which can help improving the accuracy and performance of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

prediction. Jianhui et al. proposed the LSTM-FWA model based on LSTM and FWA (FireWorks Algorithm). The model is optimized with temporal, spatial, spatio-temporal techniques respectively. Qin et al. integrated a hyperbolic model to predict PM2.5 concentrations as time series based on CNN (Convolutional Neural Network) and LSTM network. CNN network is used to extract features of input data, while LSTM network is used to consider the time dependence of air pollutants. Li et al. introduced the attention mechanism into the LSTM to capture the importance degrees of featured states at different times. The model VOLUME 9, 2021 93287 H. Chen et al.: Air Quality Prediction Based on Integrated Dual LSTM Model can predict the PM2.5 concentrations over the next 24 hours by using air quality data. Luo et al. established the BiLSTM (bidirectional long short-term memory) network, in which an EMD (empirical mode decomposition) step is introduced to reduce error accumulation in PM2.5 multi-step prediction. LSTM-based model solves the problems of gradient explosion and gradient disappearance in RNN, and has a faster learning speed. Therefore, the LSTM-based model can effectively obtain better prediction results. However, it is still difficult to obtain a high accuracy rate for data prediction because of too many factors affecting air quality changes.

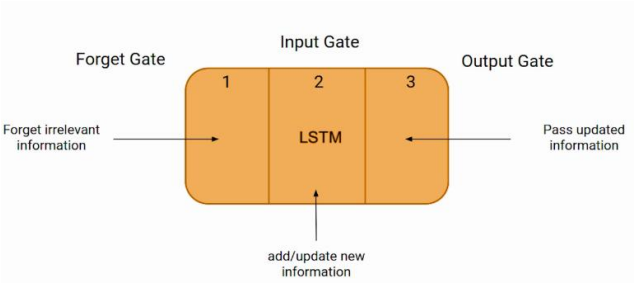
7.1 LSTM Architecture

At a high-level LSTM works very much like an RNN cell. Here is the internal functioning of the LSTM network. The LSTM consists of three parts, as shown in the image below and each part performs an individual function.



The first part chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. In the second part, the cell tries to learn new information from the input to this cell. At last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp.

These three parts of an LSTM cell are known as gates. The first part is called **Forget gate, the** second part is known as **the Input gate** and the last one is **the Output gate**

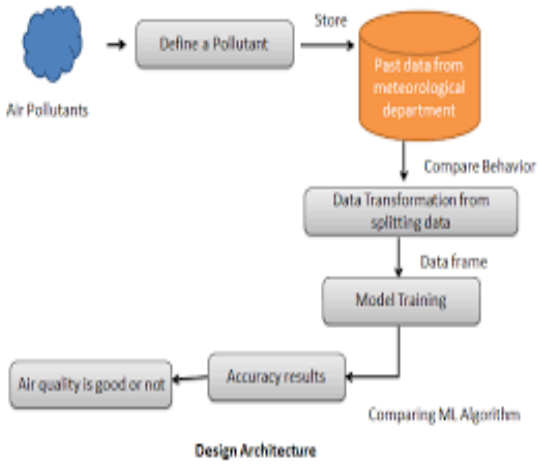


Just like a simple RNN, an LSTM also has a hidden state where H(t-1) represents the hidden state of the previous timestamp and Ht is the hidden state of the current timestamp. In addition to that LSTM also have a cell state represented by C(t-1) and C(t) for previous and current timestamp respectively.

VIII. SYSTEM ARCHITECTURE

Design is significant engineering illustration of whatever that's to be developed. Program design is a process design is e excellent option to effectively translate necessities in to a completed application product. Design creates a representation or mannequin, presents element about software information structure, architecture, interfaces and add-ons which are vital to put into effect a procedure**.**

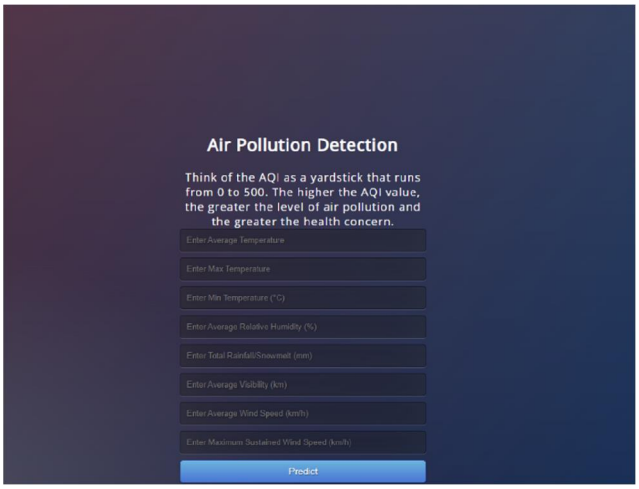
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

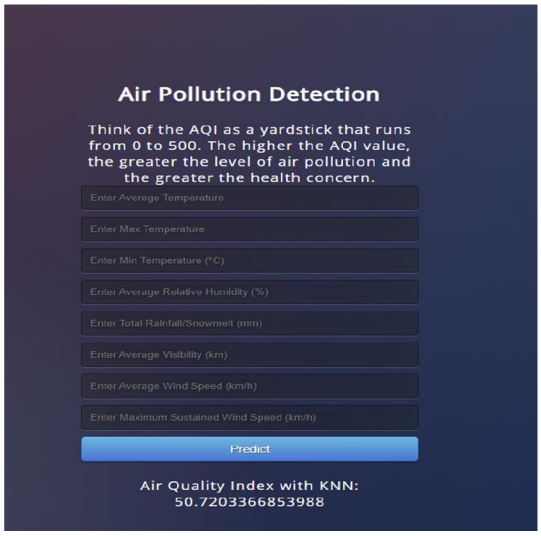


Design Architecture

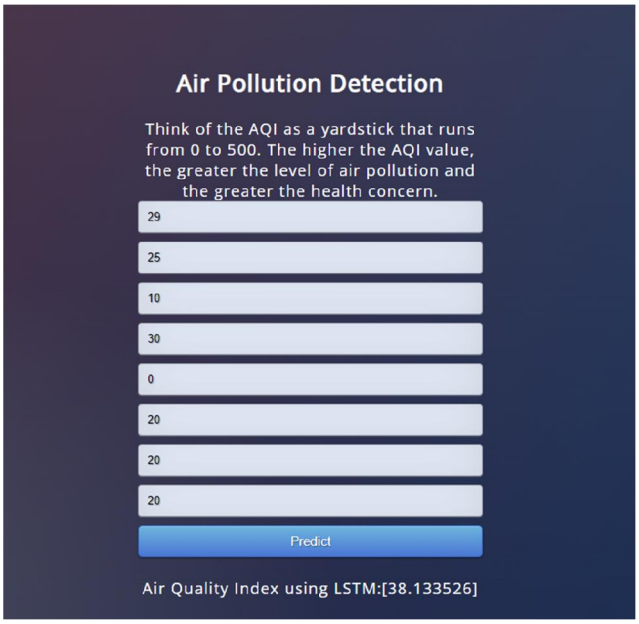
IX. RESULT/PERFORMANCE ANALYSIS

Inputs considered as values of temperature (min, max, avg), humidity, wind speed, visibility values.The outputs are as shown in the figures with the prediction of KNN and LSTM values.





IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252



X. CONCLUSION

Conclusion As atmospheric pollution concentration was not only related to the pollutant source, but also associated with many factors such as atmosphere, terrain, it is difficult to accurately predict the atmospheric pollution. Therefore, this study came up with a kind of space-time prediction of air quality situation through the KNN-LSTM hybrid model of atmospheric pollution index –AQI. Study came up with a kind of space-time prediction of air quality situation through the KNN-LSTM hybrid model of atmospheric pollution index –AQI.

The main results were concluded as follows: In this study, the air quality index was predicted with the data of provincial air pollution monitoring stations in Beijing, and an air quality index prediction method based on the KNN-LSTM hybrid algorithm model was proposed, whose effectiveness was further verified.

On the basis of time dimension, this study further took the influence of spatial distribution on the concentration of atmospheric pollutants into consideration. Through comparative experiments, it was verified that the prediction results obtained in this study were closer to the actual AQI value with a higher prediction accuracy. However, considering that the concentration of atmospheric pollutants was restricted by many factors, the prediction accuracy may be inaccurate when other factors showed a greater impact. Therefore, we suggested that the influence of meteorological, geographical and other factors should be studied in subsequent research. Using the combination of relevant data suitable prediction models should be explored to improve the prediction accuracy.

REFERENCES

[1]. Anikender Kumar, PramilaGoyal, "Forecasting of air quality in Delhi using principal component regression technique", Atmospheric Pollution Research, 2 (2011) 436-444

[2]. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, "Detection and Prediction of Air Pollution using Machine Learning Models", International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018

[3]. Heidar Malek, Armin Sorooshian, Gholamreza Goudarzi, Zeynab Baboli, Yaser Tahmasebi Birgani, Mojtaba Rahmati, "Air pollution prediction by using an artifcial neural network model", Clean Technologies and Environmental Policy, (2019) 21:1341–1352.

[4]. Nidhi Sharma , ShwetaTaneja , VaishaliSagar , Arshita Bhatt, "Forecasting air pollution load in Delhi using data analysis tools", ScienceDirect, 132 (2018) 1077– 1085.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 9, June 2022
Impact Factor: 6.252

[5]. Ziyue Guan and Richard O. Sinnot, "Prediction of Air Pollution through Machine Learning on the cloud", IEEE/ACM5th International Conference on Big Data Computing Applications and Technologies (BDCAT), 978-1-5386-5502-3/18/$31.00 2018 IEEE DOI 10.1109/BDCAT.2018.00015.

[6]. Mohamed Shakir, N. Rakesh, "Investigation on Air Pollutant Data Sets using Data Mining Tool", IEEE Xplore Part Number:CFP18OZV-ART; ISBN:978-1- 5386-1442-6.

[7]. Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu, "Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms", Applied Sciences, ISSN 2076-3417; CODEN: ASPCC7, 2019, 9, 4069; doi:10.3390/app9194069