International Journal of Recent Trends in Engineering & Research

ISSN (ONLINE) : 2455 – 1457

IMPACT FACTOR : 4.101

ICNSCET20- International Conference on New Scientific Creations in Engineering and Technology

A NOVEL SYSTEM FOR AIR POLLUTION PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHM

[1]Nagalakshmi A, [2]Samyuktha N, [3]Priskilla Angel Rani J

[1] *Student, Department of Computer Science, Anand Institute of Higher Technology, Kazhipattur,*
[2] *Student, Department of Computer Science, Anand Institute of Higher Technology, Kazhipattur,*
[3]*Assistant Professor, Department of Computer Science, Anand Institute of Higher Technology, Kazhipattur.*

----------***----------

Abstract

Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution, We are in need of implementing models which will record information about concentrations of air pollutants (SO2,NO2,etc).The deposition of this harmful gases in the air is affecting the quality of people's lives, especially in urban areas. Lately, many researchers began to use Big Data Analytics approach as there are environmental sensing networks and sensor data available.In this paper, KNN(K Nearest Neighbor) algorithm used to predict the concentration of SO2 and NO2 in the environment. Sulphur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs.Models in time series are employed to predict the SO2 and NO2 readings in nearing years or months.

Keywords **:** Machine Learning, Prediction, Air Quality, SO2 ***,NO2..***

1.INTRODUCTION

In the developing countries like India, the rapid increase in population and economic upswing in cities have lead to environmental problems such as air pollution, water pollution, noise pollution and many more. Air pollution has direct impact on humans health .There has been increased public awareness about the same in our country.Global warming, acid rains, increase in the number of asthma patients are some of the long-term consequences of air pollution. Precised air quality forecasting can reduce the effect of maximal pollution on the humans and biosphere as well. Hence, enhancing air quality forecasting is one of the prime targets for the society. Sulphur Dioxide is a gas. It is one of the major pollutants present in air.It is colorless and has a nasty, sharp smell.It combines easily with other chemicals to form harmful substances like sulphuric acid, sulfurous acid etc. Sulfur dioxide affects human health when it is breathed in. It irritates the nose, throat, and airways to cause **coughing, wheezing, shortness of breath**, or a tight feeling around the chest. The concentration of sulphur dioxide in the atmosphere can influence the **habitat suitability** for plant communities, as well as animal life. Nitrogen dioxide causes the problems such as wheezing,coughing,colds,flu and bronchitis.Nitrogen dioxide inflames the lining of the

lungs , and it can reduce immunity to lung infection.The proposed system is capable of predicting concentration of Sulphur Dioxide and Nitrogen Dioxide for forthcoming months / years..

2. RELATED WORK

In this research paper the students have forecasted the air quality of India by using machine learning algorithms to predict the air quality index(AQI) of a given area. Air quality Index is a standard measure to determine the quality of air. Concentrartion of Gases such asso2, no2,co2, rspm, spm. etc. are recorded by the agencies . These students have developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem. They improved the efficiency of the model by applying cost Estimation for predictive Problem. They say that this model is capable of successfully predicting the air quality index of a total county or any state or any bounded region provided with the historical data of pollutant concentration. The normal csv data import to MySQL.

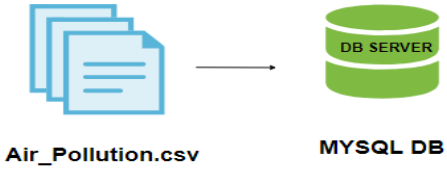

Fig 1:CSV - MYSQL

This paper implemented by using KNN (K Nearest Neighbor) algorithm. The Hadoop Tool also used to implementation. Sqoop command is used to import and export the data from MySQL to HDFS and HDFS to MySQL.

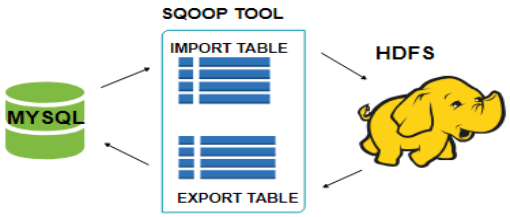

Fig 2: Data impot and export

This system has used for prediction of the pollution of next day. The system helps to predict next date pollution details based on basic parameters and analyzing pollution details and forecast future pollution. Time Series Analysis was also used for recognition of future data points and air pollution prediction. This proposed system does two important tasks (i)Analyze the data. (ii) Predict the data

MapReduce concept also discussed.Hive is used for analyze purpose

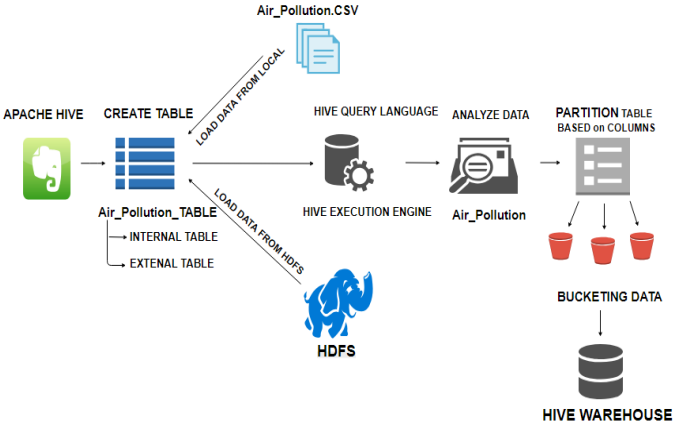

Fig 3 : Working procedure of Hive

3. DATASET

3.1 Dataset/Source: Kaggle

Structured/Unstructured data: Structured Data in CSV format.

Dataset Description: The dataset consists of around 450000 records of all the states of India.We worked only on Dataset of India.So we had 60383 records. This dataset consist of 19 attributes listed below.

1)ID

2)CITY

3)AREA

4)SEASON

5)TRAFFIC

6)CLIMATE

7)DATE

8)DAYS

9)MORNING

10)AFTERNOON

11)EVENING

12)NIGHT

13)POLLUTION

14)VEHICLES

15)TYPES

16)AVG_POL

17)SO2

18)NO2

19)OTHER

Splitting for Testing : Data Splitting was done as 80% for training and 20% for testing.

Preprocessing and Feature Selection:
We only studied and applied algorithms on the data of India.Hence, no. of rows was reduced to 60,383 and state column automatically is of no more use. All the values in pm2_5 were null values ,so we dropped the column.The ID have nothing to do with how much polluted the state. Similarly, Country is also not useful.The date is a cleaner representation of sampling_date attribute and so we will eliminate the redundancy by removing the latter. location_monitoring_station attribute is again unnecessary as it contains the location of the monitoring station which we do not need to consider for the analysis.
So, to summarize we have deleted the following features from our dataset :

Country,State,City,Place,Avg,Max,Min,Pollutants.

We have simplified the type attribute to contain only one of the three categories: industrial, residential, other.For SO2 and NO2, we replaced nan values by mean.For date, we have dropped nan values as there were only 3 null values. So after pre-processing our dataset contains 60,380 rows and 8 columns.

4. EXPLORATORY DATA ANALYSIS:

The below graph shows concentration of SO2 and NO2 over the years.

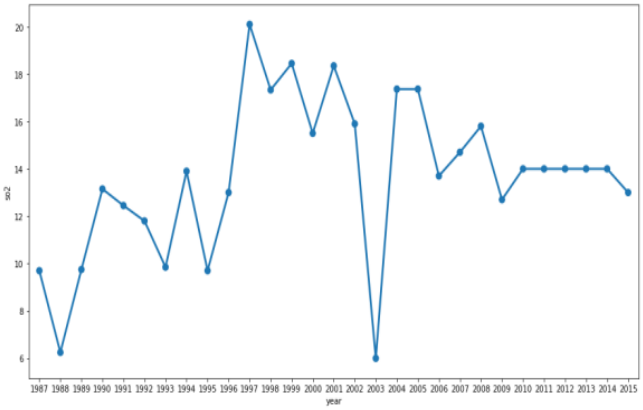

Fig 4: concentration of SO2 and NO2

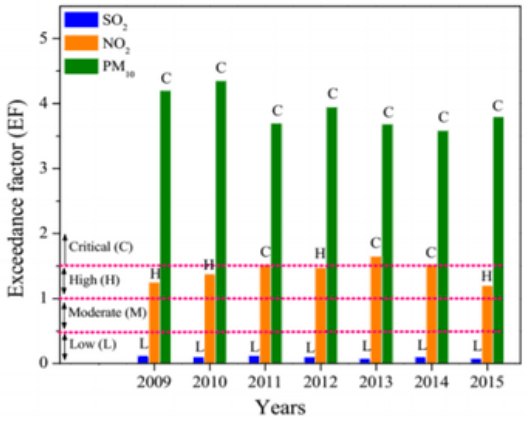

Fig 5: Pollution level based on exceedance factor (EF) of SO2 and NO2 in India

This graph shows that the amount of SO2 is highest in the industrial areas.

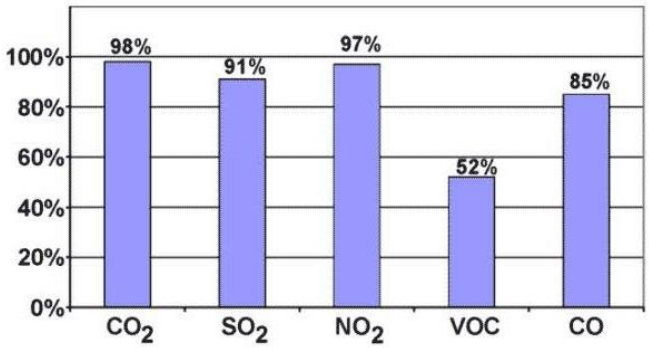

Fig 6 :Percentage of emissions from fossil fuel in US

The GUI model of air pollution prediction is given below.

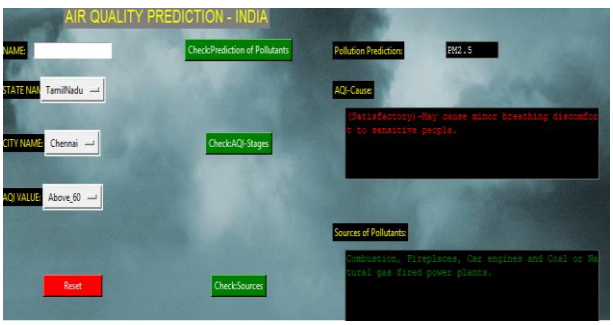

Fig 7 :Prediction of air quality

5. RESULT AND DISCUSSION:

1)AR model:(autoregressive model)

Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step. It is a very simple idea that can result in accurate forecasts on a range of time series problems.

$$yhat = b0 + b1*X1$$

Where yhat is the prediction, b0 and b1 are coefficients found by optimizing the model on training data, and X is an input value. This technique can be used on time series where input variables are taken as observations at previous time steps, called lag variables.

For example, we can predict the value for the next time step (t+1) given the observations at the last two time steps (t-1 and t-2). As a regression model, this would look as follows:

$$X(t+1) = b0 + b1*X(t-1) + b2*X(t-2)$$

2)ARIMA MODEL:

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. ARIMA is a generalization of the simpler Autoregressive Moving Average and adds the notion of integration. AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.

I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.

MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations. Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model, also called the lag order.

d: The number of times that the raw observations are differenced, also called the degree of differencing.

q: The size of the moving average window, also called the order of moving average.

6. CONCLUSION

Based on the bar plots plotted we come to the conclusion that some cities are highly polluted and need urgent attention. where concentration of SO2 is increasing, we can take measures from now to not face problems later.We used AR model and ARIMA model for predicting values of SO2,NO2. Features such as location_monitoring_station or station code were of no use as they have nothing to do with SO2 and NO2 predictions.

SO2 safe levels are as follows:

0.20 ppm (parts per million) averaged over a one hour period. 0.08 ppm averaged over a 24 hour period. 0.02 ppm averaged over a one year period.

NO2 safe levels are as follows:

0.1 ppm averaged over a one hour. NO2 should not exceed 5 ppm.

In order to predict air quality, pm2_5 is also an important attribute. The values of this must be recorded in future as this particulates are responsible for various health effects including cardiovascular effects such as cardiac arrhythmias and heart attacks, and respiratory effects such as asthma attacks and bronchitis. This model is not able to show expected output as the data is not in sequence as per date column. The same is the problem for cities. If we predict for the entire state, it won't be helpful So we will be now calculating AQI and use classification models further.

This model further, also makes us aware of the challenges in future and research needs such as pm 2.5,AQI,etc.

7. REFERENCES

[1] Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C. "Indian Air Quality Prediction And Analysis Using Machine Learning". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue)

[2] Suhasini V. Kottur , Dr. S. S. Mantha. "An Integrated Model Using Artificial Neural Network(Ann) And Kriging For Forecasting Air Pollutants Using Meteorological Data". International Journal of Advanced Research in Computer and Communication Engineering ISSN (Online) : 2278-1021 ISSN (Print) : 2319-5940 Vol. 4, Issue 1, January 2015

[3] RuchiRaturi, Dr. J.R. Prasad ."Recognition Of Future Air Quality Index Using Artificial Neural Network".International Research Journal of Engineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

[4] 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

[5] Gaganjot Kaur Kang, Jerry ZeyuGao, 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]https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

[7]https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[8] Radha Krishnan, B., Vijayan, V., Parameshwaran Pillai, T. and Sathish, T., 2019. Influence of surface roughness in turning process—an analysis using artificial neural network. Transactions of the Canadian Society for Mechanical Engineering, 43(4), pp.509-514.

[9] Krishnan, B.R., Ramesh, M., Giridharan, R., Sanjeevi, R. and Srinivasan, D., Design and Analysis of Modified Idler in Drag Chain Conveyor. International Journal of Mechanical Engineering and Technology, 9(1), pp.378-387.

[10] Krishnan, B.R., Vijayan, V. and Senthilkumar, G., 2018. Performance analysis of surface roughness modelling using soft computing approaches. Applied Mathematics and Information Sci, 12(6), pp.1209-1217.

[11] KRISHNAN, B.R. and PRASATH, K.A., 2013. Six Sigma concept and DMAIC implementation. International Journal of Business, Management & Research (IJBMR), 3(2), pp.111-114.

[12] B.Radha Krishnan., 2020. Review Of Surface Roughness Prediction In Machining Process By Using Various Parameters. International Journal of Recent Trends in Engineering & Research (IJRTER), Volume 6, Issue 1, pp.7-12.

[13] Krishnan, B.R., Sundaram, C.M. and Vembathurajesh, A., 2018. Review of Surface Roughness Prediction in Cylindrical Grinding process by using RSM and ANN. International Journal of Recent Trends in Engineering and Research, 4(12), pp.2455-1457.

