International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656

Air Quality Prediction using Machine Learning Algorithms

Pooja Bhalgat
Student
M.Sc(Big Data Analytics)
MIT-WPU, Pune, India

Sejal Pitale
Student
M.Sc(Big Data Analytics)
MIT-WPU, Pune, India

Sachin Bhoite
Assistant Professor
Computer Science
MIT-WPU, Pune, India

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,Weare 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, machine learning techniques are used to predict the concentration of so2 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 readings in nearing years or months.

Keywords: Machine Learning, Time Series, Prediction, Air Quality, SO2

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.

The proposed system is capable of predicting concentration of Sulphur 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.[1]

This paper presents an integrated model using Artificial Neural Networks and Kriging to predict the level of air pollutants at various locations in Mumbai and Navi Mumbai using past data available from meteorological department and Pollution Control Board. The proposed model is implemented and tested using MATLAB for ANN an R for Kriging and the results are presented.[2]

This system has used the Linear regression and Multilayer Perceptron (ANN)Protocol 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.[3]

This proposed system does two important 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 used 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.[4]

The major objective of this paper was to provide a snapshot of the vast research work and useful review on the current state-of-the-art on applicable big data approaches and machine learning techniques for air quality evaluation and predication.Air quality map were illustrated and visualized using data from Shenzhen, China. Artificial neural network

(ANN), Genetic Algorithm ANN Model, Random forest, decision tree, Deep belief network are the algorithms which were used and various pros and cons of the model were presented.[5]

.

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 Maharashtra.So we had 60383 records. This dataset consist of 13 attributes listed below.

1)stn_code
2)sampling_date
3) state
4) location
5) agency
6)type
7)so2
8)no2
9)rspm
10) spm
11)location_monitoring_station
12)pm2_5
13)date

Station code is a code given to each station that recorded the data,sampling date is the date when the data is recorded.state and location represents state and cities whose data is recorded and agency is the name of agency that recorded the data.Type states the type of area where the data was recorded such as industrial,residential,etc.so2,no2,rspm and spm is the amount of sulphur dioxide, nitrogen dioxide, respirable suspended particulate matter and suspended particulate matter measured respectively.date is a cleaner version of sampling_date. PM2.5 refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair.But majority of values in this column are null.

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

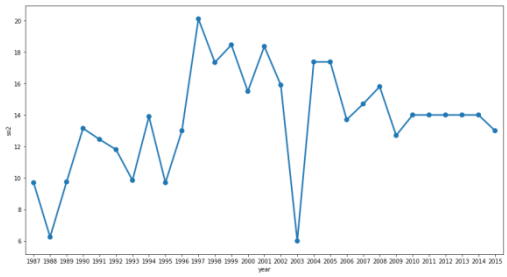
Preprocessing and Feature Selection:
We only studied and applied algorithms on the data of Maharashtra State .Hence, no. of rows was reduced to 60,383 and state column automatically is of no more use.
All the values in pm2_5 were null values ,so we dropped the column.The agency's name have nothing to do with how much polluted the state is. Similarly, stn_code 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 :

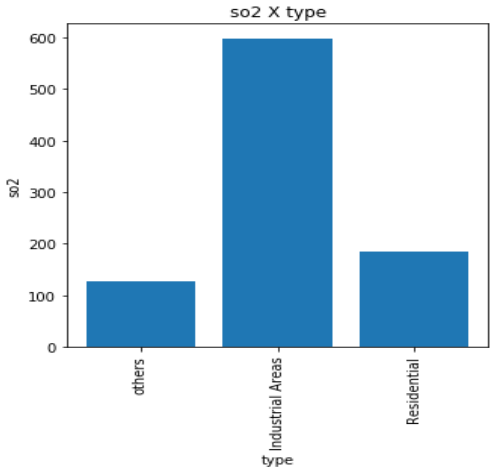
state,pm2_5,agency, stn_code, sampling_date and location_monitoring_station
We have simplified the type attribute to contain only one of the three categories: industrial, residential, other.For SO2 and NO2, we replaced nan values by mean.For date, we have dropped nan values as there were only 3 null values.
So after pre-processing our dataset contains 60,380 rows and 7 columns.

4. EXPLORATORY DATA ANALYSIS:

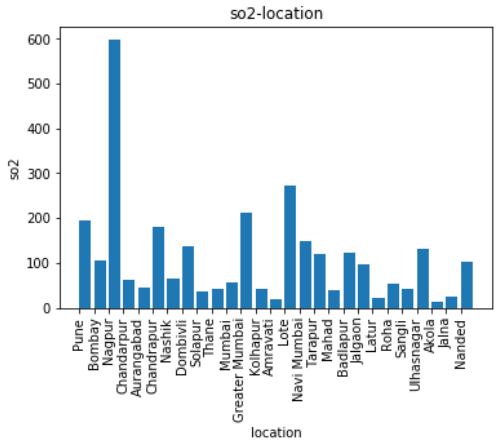
- The below graph shows concentration of so2 over the years.It was highest in the years of 1997 and 2001 and lowest in the years 1988 and 2003 .However,it is stable for the latest years.



- This graph shows that the amount of so2 is highest in the industrial areas.



- From this graph we can conclude that Nagpur has the deadliest amount of so2 as compared to other cities whereas Akole , Amravati are sparsely polluted followed by Jalna and Kolhapur.



5. RESULT AND DISCUSSION:

We are able to identify the future data points using Time Series Analysis.
Models used for the same are :

1)AR model:(autoregressive model)

Test MSE: 166.358
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)
Because the regression model uses data from the same input variable at previous time steps, it is referred to as an autoregression (regression of self).[6]

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.[7]

6. CONCLUSION

Based on the bar plots plotted we come to the conclusion that some cities are highly polluted and need urgent attention. Also for cities like Pune ,Mumbai 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. Features such as location_monitoring_station or station code were of no use as they have nothing to do with so2 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.

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 respiratoryeffects 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 wont 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 pm2.5,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/