Air Quality Prediction and Analysis using machine learning

Harsh Agarwal, Pooja Jaiswal, Kumud Jain

Deptt of Computer Science and Engineering, VIT BHOPAL University, Madhya Pradesh, India
Deptt of Computer Science and Engineering, VIT BHOPAL University, Madhya Pradesh, India
Deptt of Computer Science and Engineering, VIT BHOPAL University, Madhya Pradesh, India

Email:- harsh.agarwal2019@vitbhopal.ac.in pooja.jaiswal2019@vitbhopal.ac.in, kumud.jain2019@vitbhopal.ac.in

Abstract

Everyone knows that pollution is a major concern in modern times. There are various types of pollution wiz Water, Land, Space, Air, Sound and also we have different types of pollutants which make the environment unfit for the human race for the living.

In this paper, we are calculating the Air Quality Index of some cities of the Republic of India using machine learning algorithms. We have analyzed various attributes for AQI such as PM2.5, PM10, SO2, NOx, NH3, CO, O3, and others. Data for the study is being collected from different cities of India over a period of time. We have created the model for the same, based on the historical data obtained from the previous years. We have a total of 22 attributes for the study and a large amount of data(Which is available to us using different sites specially kaggle). The attributes for the study are as follows.

Attributes
City
Date

PM2.5
PM10
NO
NO2
NOx
NH3
CO
SO2
O3
Benzene
Toluene
Xylene
AQI
AQI_Bucket
Year
Month
StationId
StationName
State
Status
Datetime

City will be the name of the locality where the study is being conducted, at the specific date. In that study the quantity of PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene (C6H6), Toluene, Xylene are calculated. Hence the cumulative AQI and AQI_Bucket with the respective year and month with StationId and Station Name is calculated. This large amount of data is used to learn the model and the expected accuracy of the model is quite high.

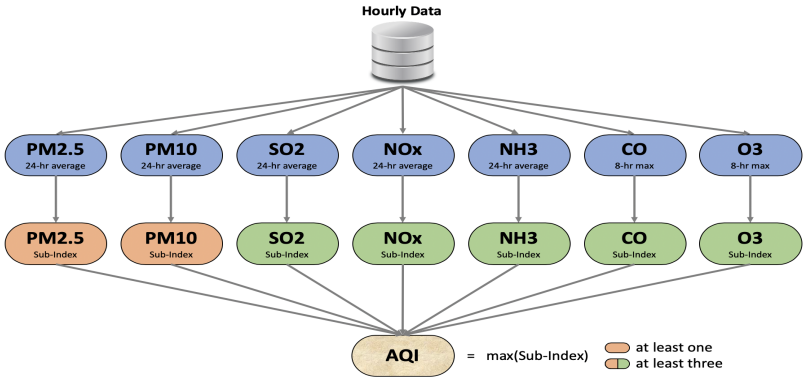
Our analysis will be helpful to predict the recent trends of Air quality in a specified region(Nation, state, or any bounded region). The analysis will be helpful to increase the overall accuracy as well as working efficiency of any organization (either governmental or non-governmental) in the field of air pollution control. The paper finds the answer to modern questions and their way to reduce them using different algorithms of Machine learning.

Keywords:- Air Quality Index, Machine Learning, AQI, Prediction

Introduction

Air Quality Index (AQI) or we can say Air Pollution Index is generally the parameter used to observe the harmfulness of the quality of the environment in a particular surrounding or a region. In the previous years, researches were made for the same by different researchers, environmentalists, and other renowned agencies but none of them is accepted worldwide.

Air pollution can be clearly defined as a global problem of the environment that generally influences the health of urbanized populations, especially the people living in metropolitan cities. Industrial nations such as China, USA, Germany, India, Japan, and others are producing a large number of pollutants such as CO2, PM 2.5, and many others. There are many pollution contributing gasses and generally, each pollution has an individual scale and indexes at different levels. Over some past years, epidemiological studies show the higher adverse effect on the health of the civilians due to the same issue. The model built will help to access the same problem with very refined accuracy.
Over the years, observations state that many cities access aqi using monitoring networks designed to measure and record the air pollution.



Identifying Design Criteria for an Ideal Air Quality prediction

The primary concept of any air quality prediction is to change all the measured particle contents to a single numerical index using a suitable particular aggregate mechanism on an individual basis. Ideally, the index is used to represent both the measured and publicly discerned ambient air quality for the time period it covers. Thus the result is air quality indices that attempt to standardize and synthesize all information regarding air pollution and allow comparisons to be taken readily and thus satisfy public demands for accurate, easy, and interpreted data.

In the design of air quality indicators, the following criteria should be used. 1. Be readily accessible by the public.

2. Include the major criteria of adulterants and their synergisms.

3. be expandable for other adulterants and comprising times.

4. Be related to National Ambient Air Quality norms used in individual businesses.

5. Avoid " surpassing" (surpassing occurs when an air pollution indicator doesn't indicate poor air quality despite the fact that attention of one or further air adulterants may have reached unacceptably grandly values).

6. Avoid " nebulosity" (nebulosity occurs when an air pollution Indicator gives false alarm despite the fact that attention of all the adulterants are within the admissible limit except one).

7. Be usable as an alert system.

8. Be grounded on valid air quality data obtained from monitoring stations that are positioned so as to represent the general air quality in the community.

Good (0–50)	Minimal Impact	Poor (201–300)	Breathing discomfort to people on prolonged exposure
Satisfactory (51–100)	Minor breathing discomfort to sensitive people	Very Poor (301–400)	Respiratory illness to the people on prolonged exposure
Moderate (101–200)	Breathing discomfort to the people with lung, heart disease, children and older adults	Severe (>401)	Respiratory effects even on healthy people

The Air Quality Prediction Model

1. **Deep System Analysis**

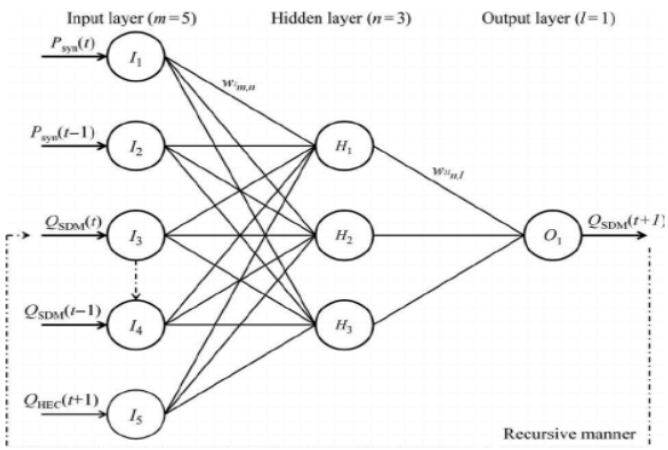
As we know Fine material Particulate material (PM2.5) is important as it impacts people's health hence it's one of their main concerns too if its level within the air reaches high. PM2.5 refers to the little particles staying in the air that scale black visibility and air to seem hazy to naked eyes when its level gets elevated.
Here in the system we proposed, we calculated the air index quality of all the pollutants present in the atmosphere using AQI formulas to know the air quality in India using gradient descent and Box-Plot analysis. The system proposed can also be utilized in the upcoming years to predict the air index using present AQI values.

2. **Back Propagation**

The technique used in fake neural networks which is generally called Back Propagation is used to figure out
an inclination that is required in the count of the loads to be utilized in the network. Backpropagation is a good choice for "the retrogressive proliferation of mistakes". Generally, people do use 2 types of backpropagation.

- Static back-propagation
- Recurrent back-propagation.

One good part of back-propagation is that It has almost zero or minimal effect on neural networks.
Some common work which we focused on while applying backpropagation is first, it is very much inclined on the input data and hence a lot of effort is needed. Secondly, we found that it was quite sensitive to the missing and noisy data, this is one major motivation to reduce the noise from the model.



EXPERIMENTAL ANALYSIS

A. Data Sources

To predict the regional air quality index, we must compute the concentration of the pollutants of all the gases that are present in the air, which actually holds all the data regarding the pollutants that pollute the cities every year.

The AQI formulae are applied to calculate the AQI by using a linear regression algorithm for a particular year. During the computation, there are various databases that have been imported and the null values have been set to the infinite data. As an outcome, the predicted values and the actual values have been represented below using scatter plot and linear analysis in order to remove the outliners.

B. Pre Processing of data:The data cleaning

In the dataset used, the outliners are basically of faulty sensor or transmission errors, these errors have a huge variation in values when compared to normal vivid results. As we know that the standard range of these pollutants occur on a particular area hence to remove these outliners from the data we had used boundary value analysis. By the help of BVA we had planted the upper quartile range and lower quartile range for the dataset we took.

	No of missing values	% of missing values
Xylene	18109	61.320000
PM10	11140	37.720000
NH3	10328	34.970000
Toluene	8041	27.230000
Benzene	5623	19.040000
AQI	4681	15.850000
AQI_Bucket	4681	15.850000
PM2.5	4598	15.570000
NOx	4185	14.170000
O3	4022	13.620000
SO2	3854	13.050000
NO2	3585	12.140000
NO	3582	12.130000
CO	2059	6.970000

The above image shows us how many missing values present in our dataset.

C.AQI Simulation and calculation

We have acquainted the database with a number of various column of sensor data with vivid places in India. We have collected the average reading of ambivert air quality with respect to parameters of air quality like Sulphur dioxide(SO2), Nitrogen dioxide(NO2) , Respirable suspended particulate material(RSPM) and Suspended Particulate Material (SPM). Also, the database used also consist of many noisy data itself since a few data stations have been shifted or closed at the period when they were marked as NAN or were not available. Hence, we have pre-processed the data in order to remove these outliners only. Each of the individual pollutant indices, gives the relationship between the pollutants concentration and their individual indices respectively. Hereby, the example images of individual AQI calculations have been given.

```
## PM2.5 Sub-Index calculation
def get_PM25_subindex(x):
    if x <= 30:
        return x * 50 / 30
    elif x <= 60:
        return 50 + (x - 30) * 50 / 30
    elif x <= 90:
        return 100 + (x - 60) * 100 / 30
    elif x <= 120:
        return 200 + (x - 90) * 100 / 30
    elif x <= 250:
        return 300 + (x - 120) * 100 / 130
    elif x > 250:
        return 400 + (x - 250) * 100 / 130
    else:
        return 0

df["PM2.5_SubIndex"] = df["PM2.5_24hr_avg"].apply(lambda x: get_PM25_subindex
(x))
```

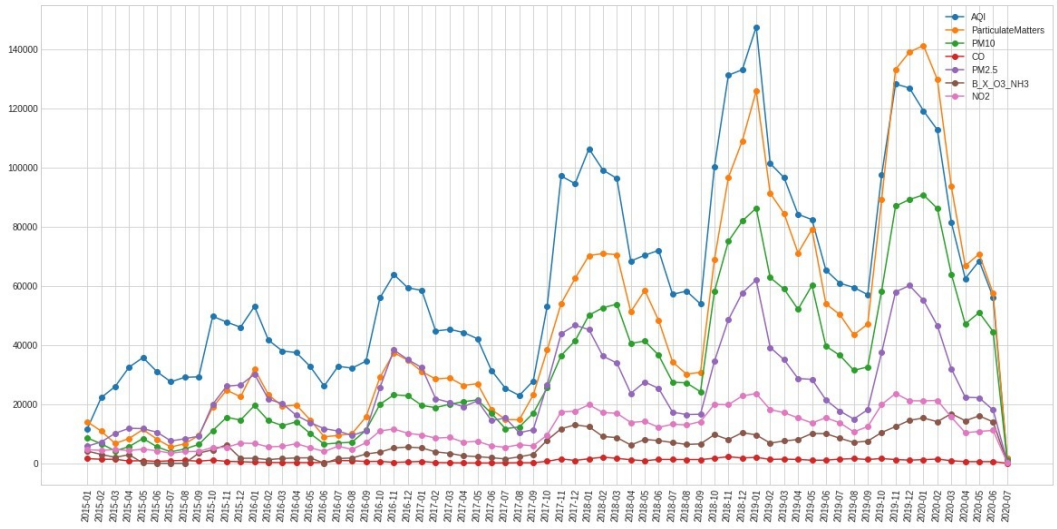
```
## AQI bucketing
def get_AQI_bucket(x):
    if x <= 50:
        return "Good"
    elif x <= 100:
        return "Satisfactory"
    elif x <= 200:
        return "Moderate"
    elif x <= 300:
        return "Poor"
    elif x <= 400:
        return "Very Poor"
    elif x > 400:
        return "Severe"
    else:
        return np.NaN
```

```
df["Checks"] = (df["PM2.5_SubIndex"] > 0).astype(int) + \
                (df["PM10_SubIndex"] > 0).astype(int) + \
                (df["SO2_SubIndex"] > 0).astype(int) + \
                (df["NOx_SubIndex"] > 0).astype(int) + \
                (df["NH3_SubIndex"] > 0).astype(int) + \
                (df["CO_SubIndex"] > 0).astype(int) + \
                (df["O3_SubIndex"] > 0).astype(int)

df["AQI_calculated"] = round(df[["PM2.5_SubIndex", "PM10_SubIndex", "SO2_SubIn
dex", "NOx_SubIndex",
                                 "NH3_SubIndex", "CO_SubIndex", "O3_SubInde
x"]].max(axis = 1))
df.loc[df["PM2.5_SubIndex"] + df["PM10_SubIndex"] <= 0, "AQI_calculated"] = n
p.NaN
df.loc[df.Checks < 3, "AQI_calculated"] = np.NaN

df["AQI_bucket_calculated"] = df["AQI_calculated"].apply(lambda x: get_AQI_buc
ket(x))
df[~df.AQI_calculated.isna()].head(13)
```

Result analysis

Line plot is used most commonly in the exploratory data analysis for predicting the results. We will be using the same so that the data can be easily shown and portrayed.

The same is used for the data prediction in our work. The below graph is a combination of clusters of graphs that indicate and reflects the data.



The data shows the comparison between all twenty-two variables with Air quality prediction.

Conclusion and future scope

Since the accuracy for predicting the data is around 93 percent, so it will calculate almost all the data present in the modern era for a particular region. The data and the analysis will be helpful to alert any region in the country if the air quality decreases and hence appropriate actions can be taken on the same. Also, it is worth noticing that the data is taken from Kaggle, hence we can say that the model uses real-time data for the analysis. One can incorporate app and web interfaces in the model to make it user-friendly, For security, the use of different communications channels can be utilized, and cyber security protocols with the model can be utilized. This model can help the future generation for solving the problem of air quality in the environment to a very large extend.

References:-

[1] Bezuglaya, E.Y., Shchutskaya, A.B., Smirnova, I.V. (1993) Air Pollution Index and Interpretation of Measurements of Toxic Pollutant Concentrations. Atmospheric Environment 27, 773-779.
[2] Bishoi, B., Prakash, A., Jain, V.K. (2009) A comparative study of air quality index based on factor analysis and US-EPA methods for an urban environment. Aerosol Air Quality Research 9(1), 1-17.
[3] Bruce, N., Perez-Padilla, R., Albalak, R. (2000) Indoor air pollution in developing countries: a major environmental and public health challenge. Bulletin of World Health Organisation 78(9), 1078-1092.
[4] Cairncros, E.K., John, J., Zunckel, M. (2007) A Novel Air Pollution Index Based on the Relative Risk of Daily Mortality Associated with Short-term Exposure to Common Air Pollutants. Atmospheric Environment 41, 8442-8454.
[5] Cannistraro, G., Ponterio, L. (2009) Analysis of Air Quality in the Outdoor Environment of the City of Messina by an Application of the Pollution Index Method. International Journal of Civil and Environment Engineering 1, 4.
[6] Cheng, W.L., Kuo, Y.C., Lin, P.L, Chang, K.H., Chen, Y.S., Lin, T.M., Huang, R. (2004) Revised air quality index derived from an entropy function. Atmospheric Environment 38, 383-391.
[7] Dunteman, G.N. (1994) In Factor Analysis and Related Techniques. Vol. 5, Lewis-Beck, M.S. (Ed.), Sage Publications, London, 157.
[8] Gorai, A.K., Kanchan, Upadhyay, A., Goyal, P. (2014) Design of fuzzy synthetic evaluation model for air quality assessment. Environment Systems and Decisions 34, 456-469. doi 10.1007/s10669-014-9505-6.
[9] Gorai, A.K., Tuluri, F., Tchounwou, P.B. (2014) A GIS Based Approach for Assessing the Association between Air Pollution and Asthma in New York State, USA. International Journal Environmental Research and Public Health 11(5), 4845-4869. doi:10.3390/ijerph110504845.
[10] Harman, H.H. (1968). Modern Factor Analysis, 2nd Ed., Revised. University of Chicago Press, Chicago.
[11] Hämekoski, K. (1998). The Use of a Simple Air Quality Index in the Helsinki Area, Finland. Environment Management 22(4), 517-520.

[12] Jain, R.K, Urban, L.V., Stacey, G.S. (1977) Environmental Impact analysis. Van Nustrand Reinhold, New York, 170-187.
[13] Johnston, R.J. (1978) Multivariate Statistical Analysis in Geography, Longman, New York.
[14] Kumar, A., Goyal, P. (2013) Forecasting of Air Quality Index in Delhi Using Neural Network Based on Principal Component Analysis. Pure and Applied Geophysics 170, 711-722. doi: 10.1007/s00024-012-0583-4.
[15] Kyrkilis, G., Chaloulakou, A., Kassomenos, P.A. (2007) Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration: Relation to potential health effects. Environment International 33, 670-676.
[16] Leeuw, de F., Mol, W. (2005) Air quality and air quality indices: a world apart? European Topic Centre on Air and Climate Change, Technical paper 2005/5.
[17]Lohani, B.N. (1984). Environmental Quality Management, South Asian Publishers, New Delhi.
[18] Mandal, T., Gorai, A.K., Pathak, G. (2012) Development of fuzzy air quality index using soft computing approach. Environmental Monitoring and Assessment 184, 6187-6196. doi: 10.1007/s10661-011-2412-0.
[19] Maynard, R.L., Coster, S.M. (1999) Informing the Public about Air Pollution. In Air Pollution and Health, eds. S.T. Holgate, J.M. Samet, H.S. Koren, and R.L.
[20] Maynard, pp. 1019-1033. San Diego, CA: Academic Press.
[21] Murena, F. (2004) Measuring Air Quality over Large Urban Areas: Development and Application of an Air Pollution Index at the Urban Area of Naples.
[22] Ott, W.R., Hunt, W.F. Jr. (1976) A Quantitative Evaluation of the Pollutant Standards Index. Journal of the Air Pollution Control Association 26, 1050-1054.
[23] Ott, W.R., Thom, G.C. (1976) A Critical Review of Air Pollution Index Systems in the United States and Canada. Journal of the Air Pollution Control Association 26, 460-470. https://doi.org/10.1080/00022470.1976.10470272
[24] Pyta, H. (2008) Classification of air quality based on factors of relative risk of mortality increase. Environment Protection Engineering 34(4), 111-117.
[25] Qian, Z., Chapman, R.S., Hu, W., Wei, F., Korn, L.R., Zhang, J. (2004) Using air pollution based community cluster to explore air pollution health effects in children. Environment International 30, 611-620. https://doi.org/10.1016/j.envint.2003.11.003

[26] Radojevic, M., Hassan, H. (1999) Air quality in Brunei Darussalam During the 1998 Haze Episode. Atmospheric Environment 33, 3651-3658. https://doi.org/10.1016/S1352-2310(99)00118-1
[27] Schwartz, J. (1994) Air pollution and hospital admissions for the elderly in Birmingham, Alabama. American Journal of Epidemiology 139(6), 589-598.
[28] Shenfeld, L. (1970) Note on Ontario's Air Pollution Index and Alert System. Journal of the Air Pollution Control Association 20, 612. https://doi.org/10.1080/00022470.1970.10469451
[29] icard, P., Lesne, O., Alexandre, N., Mangin, A., Collomp, R. (2011) Air quality trends and potential health effects - Development of an aggregate risk index. Atmospheric Environment 45, 1145-1153. https://doi.org/10.1016/j.atmosenv.2010.12.052
[30] Singh, G. (2006). An index to measure depreciation in air quality in some coal mining areas of Korba industrial belt of Chhattisgarh, India. Environmental Monitoring and Assessment 122, 309-317. https://doi.org/10.1007/s10661-005-9182-5