INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING

Mrs. A. Gnana Soundari
MTech, (PhD)
Associate Professor
Jeppiaar Engineering College

Mrs. J. Gnana Jeslin M.E,
(PhD)
Assistant Professor
Jeppiaar Engineering College

Akshaya A.C
Student,
Jeppiaar Engineering College

Abstract---We forecast the air quality of India by using machine learning to predict the air quality index of a given area. Air quality index of India is a standard measure used to indicate the pollutant (so2, no2, rspm, spm. etc.) levels over a period. We developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as aGradient decent boosted multivariable regression problem. we improve the efficiency of the model by applying cost Estimation for our predictive Problem. Our model will be capable for 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. In our model by implementing the proposed parameter-reducing formulations, we achieved better performance than the standard regression models. our model has 96% accuracy on predicting the current available dataset on predicting the air quality index of whole India, also we use AHP MCDM technique to find of order of preference by similarity to ideal solution.

Keywords—AQI,dataset,preprocessing, outliers, BVA,prediction

I.INTRODUCTION

As the largest growing industrial nation, India is producing record amount of pollutants specifically Co2, pm2.5 etc and other harmful aerial contaminants. Air quality of a particular state or a country is a measure on the effect of pollutants on the respected regions, as per the Indian air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicates the levels of major pollutants on the atmosphere. There are various atmospheric gases which causes pollution on our environment. Each pollution has individual index and scales at different levels. The major pollutants Such as (no2, so2, rspm, spm) indexes AQI is acquired, with this individual AQI, the data can be categorized based on the limits. We collected the data from the Indian government database, which contains pollutant concentration occurring at various places across India. We start by calculating the individual index of the pollutant for every available datapoints and find their respective AQI for the region. We have designed a model to predict the air quality index of every available data points in the dataset, our model is capable of forecasting the air quality of India in any given area. By predicting the air quality index, we can backtrack the major pollution causing pollutant and the location affected seriously by the pollutant across India. With this forecasting model, various knowledge about the data are extracted using various techniques to obtain heavily affected regions on a particular region(cluster). This give more information and knowledge about the cause and seniority of the pollutants.

II.AIR QUALITY INDEX PREDICTION MODEL

A. SYSTEM ANALYSIS

Fine material (PM2.5) could be a important one as a result of it's a giant concern to people's health once its level within the air is comparatively high. PM2.5 refers to little particles within the air that scale back visibility and cause the air to look hazy once levels are elevated. But in the proposed system we calculate the air quality index of all the pollutants using the AQI formulae to know the air quality level in a particular city using gradient descent and Box-Plot analysis. In the proposed

system the air quality index of the upcoming years can be predicted using the present AQI values.

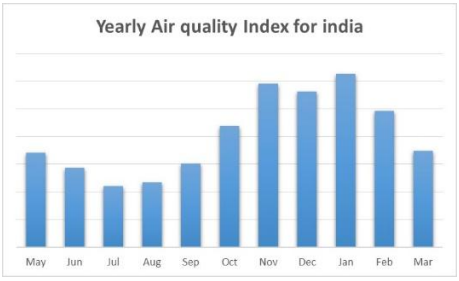


Figure 1 Air quality index

B. BACK PROPAGATION

Back propagation is a technique utilized in fake neural systems to figure an inclination that is required in the count of the loads to be utilized in the network. Back propagation is shorthand for "the retrogressive proliferation of mistakes," since a blunder is processed at the yield and appropriated in reverse all through the system's layers. It is regularly used prepare profound neural networks.

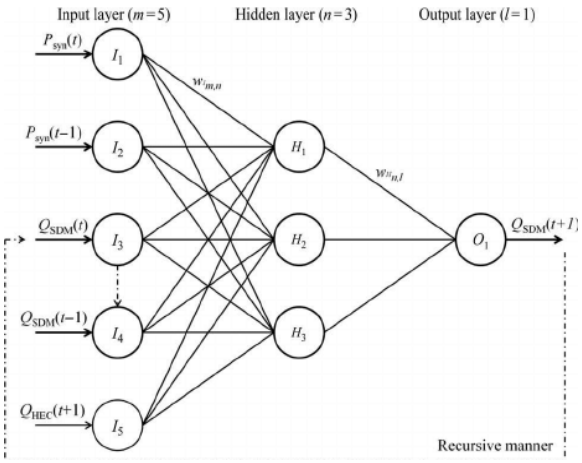


Figure 2 Neural networks

Back spread is a speculation of the delta guideline to multi-layered feed forward systems, made conceivable by utilizing the chain principle to iteratively register angles for each layer. It is firmly identified with the Gauss– Newton calculation and is a piece of proceeding with research in neural back spread

III.EXPERIMENTAL ANALYSIS

A.DATA SOURCES

To predict the air quality index of a particular region , we need the pollutant concentration of all the gases which will be available in the cpcb.nic.in website, which holds all the data that pollutes the cities every year. The AQI formulae will be applied in order to calculate the AQI by using the linear regression algorithm for a particular year. Several datasets will be imported inside the directory and null values will be set to the infinite data. The predicted and actual values will be represented using the Box-Plot analysis in order to remove the outliers.

B.PRE-PROCESSING THE DATA

In this dataset the outliers are mainly of faulty sensor or transmission errors, these errors have huge variation than the normal valid results. We know the standard range of pollutants occurs on a particular areaso to remove the outliers from the data we use boundary value analysis. By using BVA we found the upper quartile range and lower quartile range of a given data.

C.AQI SIMULATION AND CALCULATION

We acquired the dataset with various columns of sensor data from various places in India. we have

the average readings of ambient air quality with respect to air quality parameters, like Sulphur dioxide (So2), Nitrogen dioxide (No2), Respirable Suspended Particulate Matter (RSPM) and Suspended Particulate Matter (SPM). Data acquired from the source has more noisy data since few of the data from the stations have been shifted or closed the period were marked as NAN or not available.so we have to pre-process the data in order to remove the outliers.Each individual pollutant indexes, gives the relationship between the pollutant concentration and their corresponding individual index.Figure 3 shows an example of the individual AQI calculation of SO_2

```
In [3]:
#Function to calculate so2 individual pollutant index(si)
def calculate_si(so2):
    si=0
    if (so2<=40):
     si= so2*(50/40)
    if (so2>40 and so2<=80):
     si= 50+(so2-40)*(50/40)
    if (so2>80 and so2<=380):
     si= 100+(so2-80)*(100/300)
    if (so2>380 and so2<=800):
     si= 200+(so2-380)*(100/800)
    if (so2>800 and so2<=1600):
     si= 300+(so2-800)*(100/800)
    if (so2>1600):
     si= 400+(so2-1600)*(100/800)
    return si
data['si']=data['so2'].apply(calculate_si)
df= data[['so2','si']]
df.head()
```

Out[3]:

	so2	si
0	4.8	6.000
1	3.1	3.875
2	6.2	7.750
3	6.3	7.875
4	4.7	5.875

Figure 3 Calculation of SO_2

The air quality index of a particular data point is the aggregate of maximum indexed pollutant on that particular area. That pollutants maxsub index is taken as the air quality index of that particular location. Figure 4 shows the mean AQI calculation of all the gases

```
def calculate_aqi(si,ni,spi,rpi):
    aqi=0
    if(si>ni and si>spi and si>rpi):
     aqi=si
    if(spi>si and spi>ni and spi>rpi):
     aqi=spi
    if(ni>si and ni>spi and ni>rpi):
     aqi=ni
    if(rpi>si and rpi>ni and rpi>spi):
     aqi=rpi
    return aqi
```

Figure 4 AQI Calculation

Out[7]:

	sampling_date	state	si	ni	rpi	spi	AQI
0	February - M021990	Andhra Pradesh	6.000	21.750	0.0	0.0	21.750
1	February - M021990	Andhra Pradesh	3.875	8.750	0.0	0.0	8.750
2	February - M021990	Andhra Pradesh	7.750	35.625	0.0	0.0	35.625
3	March - M031990	Andhra Pradesh	7.875	18.375	0.0	0.0	18.375
4	March - M031990	Andhra Pradesh	5.875	9.375	0.0	0.0	9.375

Figure 5 Mean AQI

In this graph AQI is the average value of AQI of each year across India.

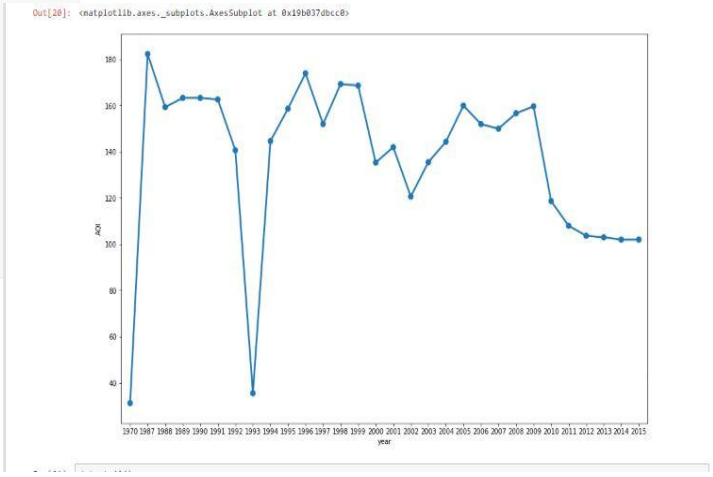


Figure 6 Graph between average AQI and sample data

D.PREDICTION OF AIR QUALITY INDEX

Using Naïve Forecast approach**,**we spitted the dataset into two parts of first 75% and rest 25% data into test and train datasets to identify the huge seasonal variations and trend.

We calculated the moving average of our datapoints and plotted the moving average. We identified the moving average varies one the year (2010-2011) i.e. before 2010 there are variations at x minimum and x maximum and after 2011 the variations are y minimum and y maximum.

Plotted the graph of train and test dataset with their moving average and analyzed the moving average. Figure 7 shows the moving average graph.

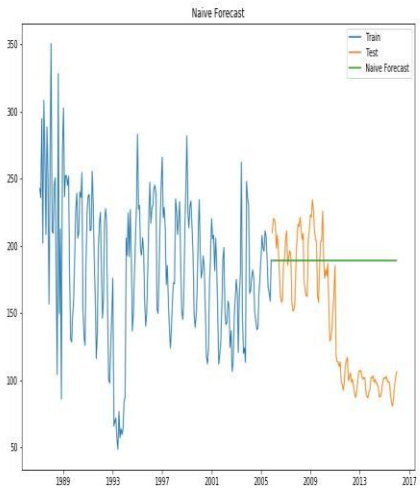


Figure 7 Moving average graph

E.RESULTS ANALYSIS

Box plot is one of common graphical systems utilized in EDA.A crate plot or boxplot is a helpful method for graphically portraying gatherings of numerical information through their quartiles. Box plots may likewise have lines broadening vertically from the containers(bristles) demonstrating inconstancy outside the upper and lower quartiles, henceforth the terms box-and-hair plot and box-and-stubblegraph. Exceptions might be plotted as individual focuses.

Box Plot gives fundamental data about a dispersion. It graphically delineates a gathering of numerical information as indicated.

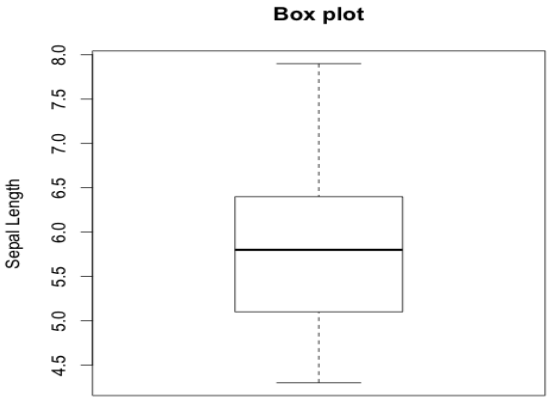


Figure 8 Box-Plot analysis

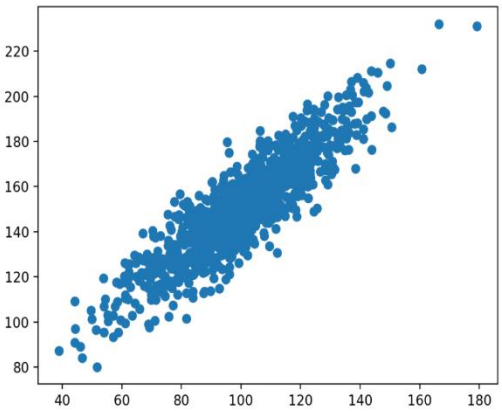
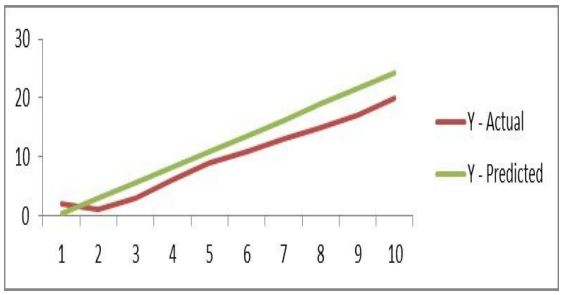


Figure 9 Testing the dataset

By this data analysis we came to know that there are seasonal variations and trend, in order to reduce these metrics, we resample the data month wise to predict it month wise. By resampling the data, we can reduce the outlier more efficiently than raw data. After removing the outlier's linear regression is applied to the filtered data and to fit the trent line on the data points gradient descent hyper parameters are used to optimize the model.

LINEAR REGRESSION

While doing straight relapse our goal is to fit a line through the dissemination which is closest to the majority of the focuses. Subsequently lessening the separation (mistake term) of information focuses from the fitted line.



F

Figure 10 Linear regression graph

Y=mx +c denotes the equation of regression line

GRADIENT BOOST ALGORITHM

The principle issue influenced by individuals is air contamination since air contains numerous substances which might be made by manmade or regular procedure. The Air substances present most organic atoms, points of interest and perilous material into the air. **Boosting Algorithm** is a victor among the most prevalent learning insights showed over the most recent twenty years.

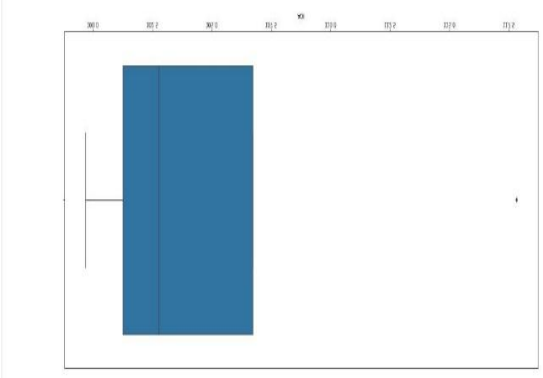


Figure 11 Outlier removal using BPA

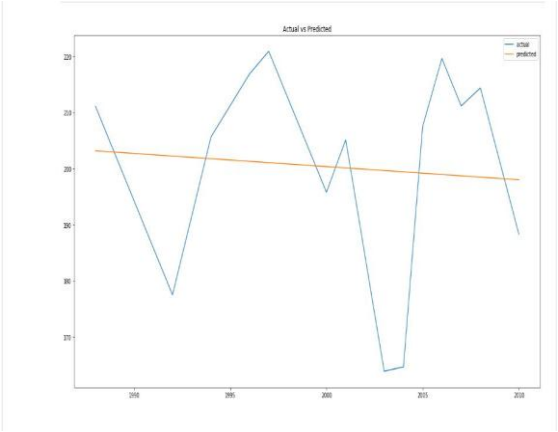


Figure 12 Actual and predicted values

CONCLUSION AND FUTURE ENHANCEMENTS

Since our model is capable of predicting the current data with 95% accuracy it will successfully predict the upcoming air quality index of any particular data within a given region. With this model we can forecast the AQI and alert the respected region of the country also it a progressive learning model it is capable of tracing back to the particular location needed attention provided the time series data of every possible region needed attention.The air quality information utilized in this paper originates from the china air quality checking and investigation stage, and incorporates the normal every day fine particulate issue (PM2.5), inhalable particulate issue (PM10), ozone (O3), CO, SO2, NO2 fixation and air quality record(AQI).The essential perspectives that should be viewed as with regards to guaging of the poison focus are its different sources alongside the components that impact its fixation.

REFERENCES

[1] Dragomir, Elia Georgiana. "Air quality index prediction using K-nearest neighbor technique no. 1 (2010): 103-108.

[2] Carbajal-Hernández, José Juan "Assessment and prediction of air quality using fuzzy logic and autoregressive models." *Atmospheric Environment* 60 (2012): 37-50.

[3] Kumar, Anikender and P. Goyal, " Forcasting of daily air quality index in Delhi", Science of th Total Environment 409, no. 24(2011): 5517-5523..

[4] Singh Kunwar P., et al. "Linear and nonlinear modelling approaches for urban air quality prediction, " Science of the Total Environment 426(2012):244-255.

[5] Sivacoumar R, et al, " Air pollution modelling for an industrial complex and model performance evaluation ", Environmental Pollution 111.3 (2001) : 471-477

[6] Gokhale sharad and Namita Raokhande, "Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period", Science of the total environment 394.1(2008): 9-24.

[7] Bhanarkar, A. D., et al, "Assessment of contribution of SO2 and NO2 from different sources in Jamshedpur region, India, "Atmospheric Environment 39.40(2005):7745-India." Atmospheric Environment 39.40 (2005): 7745-7760.

[8] Singh Kunwar P., Shikha Gupta and Premanjali Rai, " Identifying pollution sources and prediction urban air quality using ensemble learning methods", Atmospheric environment80 (2013): 426-437.

[9] Wang Jun, and Sundar A. Christopher, "Intercomparison between satellite derived aerosol optical thickness and PM2. 5 Mass: Impliances for air quality studies", Geophysical research letters30.21(2003).

[10] Sharma M E A McBean and U.Ghosh, "Prediction of atmospheric sulphate deposition at sensitive receptors in northern India", Atmospheric Environment 29.16(1995): 2157-2162.

[11] Russo Ana Frank Raischel and Pedro G.Lind, "Air quality prediction using optimal neural networks with stochastic variables", Atmospheric Environment 79(2013): 822-830.

[12] Challa Venkara Srinivas et al ," Data Assimilation and performance of Wrf for Air Quality Modeling in Mississippi Gulf Coastal Region "

[13] Hutchison Keith D., Solar Smith and Shazia J. Faruqui, " Correlating MODIS aerosol optical thickness data with ground-based PM2.5 observations across Texas for use in a real time air quality prediction system, " Atmospheric Environment 39.37(2005) :7190 – 7203

[14] Wang Z et al , " A nested air quality prediction modelling system for urban and regional scales : Application for high high-ozone episode in Taiwan " Water, Air and Soil Pollution130.1-4(2001):391-396

[15] Nallakaruppan, M. K., and U. Senthil Kumaran. "Quick fix for obstacles emerging in management recruitment measure using IOT-based candidate selection." *Service Oriented Computing and Applications* 12.3-4 (2018): 275-284.

[16] Nallakaruppan, M. K., and Harun SurejIlango. "Location Aware Climate Sensing and Real Time Data Analysis." *Computing and Communication Technologies (WCCCT), 2017 World Congress on*. IEEE, 2017.