Prediction of Air Pollution by using Machine Learning Algorithm

Jayant Kumar Singh
Galgotias University, Computer Science
Plot No. 2, Yamuna Expy Sector 17A, Greater Noida, Uttar Pradesh 203201
Jayanticfai87@gmail.com

Amit Kumar Goel
Galgotias University, Computer Science
Plot No. 2, Yamuna Expy Sector 17A, Greater Noida, Uttar Pradesh 203201
amit.goel@galgotiasuniversity.edu.in

Abstract— controlling and defensive the higher air greatness has gotten one in everything about first imperative occasions in different creating and metropolitan districts at the present. The greatness of air is adversely contacting collectible to the different styles of tainting influenced through the transportation, power, powers consumptions, and so forth. In our country population is a big problem as day by day population is increasing, so the rapid increasing in population and economic upswing is leading environment problems in city like air pollution, water pollution etc. In some of air pollution and air pollution is direct impact on human body. As we know that major pollutants are arising from Nitrogen Oxide, Carbon Monoxide & Particulate matter (PM), SO2 etc. Carbon Monoxide is arising due to the deficient Oxidization of propellant like as petroleum, gas, etc. nitrogen oxide (NO) is arising due to the ignition of thermal fuel;

Sulphur Dioxide(So2) is major spread in air, So2 is a gas which is present more pollutants in air, it's affect more in human body. the predominance of air is overstated by multi-dimensional impacts containing spot, time and vague boundaries. The goal of this improvement is to take a gander at the AI basically based ways for air quality expectation. In this paper we will predict of air pollution by using machine learning algorithm.

Keywords— Air Quality Index (AQI), Linear Regression, Python, SO2. Jupyter Notebook

I. Introduction

The Environment describe about the thing which is everything happening in encircles . the Environment is polluted by human daily activities which include like air pollution, noise pollution. If humidity is increasing more than automatically environment is going more hotter. Major cause of increasing pollution is increasing day by day transport and industries . there are 75 % NO or other gas like CO, SO2 and other particle is exist in environment.. The expanding scene, vehicles and creations square measure harming all the air at a feared rate.

Therefore, we have taken some attributes data like vehicles no., Pollutants attributes for prediction of pollution in specific zone of Delhi

A. *Literature reviews/ Comaparitive Study*

The Air Pollution Forecasting System: Air Quality Index (AQI) is a record that gives the public the degree of contamination related with its wellbeing impacts. The AQI centers around the different wellbeing impacts that individuals may encounter dependent fair and square and long stretches of introduction to the poison concentration. The AQI values are not quite the same as nation to nation dependent on the air quality norm of the country. The higher the AQI level more noteworthy is the danger of wellbeing related problems.

The by and large point of this venture is to make a student calculation that will have the option to foresee the hourly contamination focus. Additionally, an Android application will be built up that will provide the clients about the constant contamination convergence of PM2.5 alongside the hourly forecasted value of the toxin fixation from the student calculation. The Android application will also recommend data of the less dirtied[1].

II. Design

Linear Regression is basically use for predicting the real values data y using continuous parameter.

I have categorized this project in various steps and each step introduce different component or module each step define sequence and coincide to predict the AQI of a selected region and month.

III. System Design

This venture has been arranged into various advances. Each progression has an alternate segment or module answerable for at least one undertakings to be refined or executed. These means happen in succession and in synchronization to anticipate the AQI of a given district with the most elevated conceivable accuracy. With the help of this framework the client can be get particular district, months and yield details of anticipated AQI of that month. So when we will give input like location, date and months we will get predicted data from this model. for providing attributes data in system algorithm client have to set a determine area, month and time then accordingly yield value anticipated AQI for that month.. [6]

A. Linear Regression

- Linear Regression is basically use for predicting the real values data y using continuous parameter.
- It's use in several areas like financing, economics, zoology etc..."

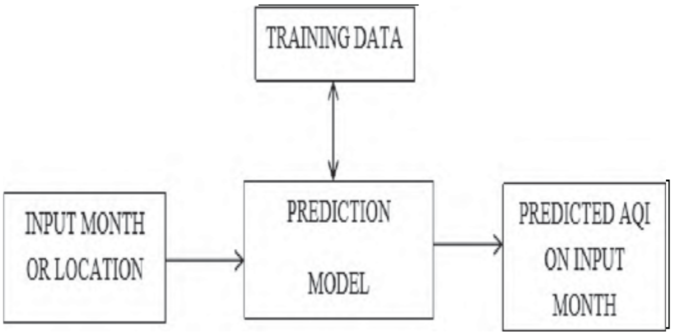


Fig 1: High Level Designs System

B. Architecture of Component Design

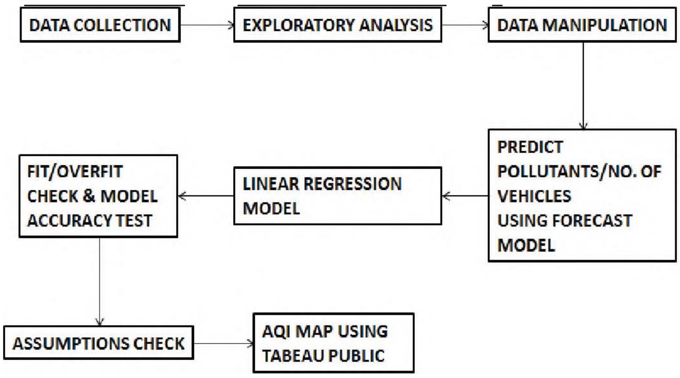

DATA COLLECTION
EXPLORATORY ANALYSIS
DATA MANIPULATION
PREDICT POLLUTANTS/NO. OF VEHICLES USING FORECAST MODEL
LINEAR REGRESSION MODEL
FIT/OVERFIT CHECK & MODEL ACCURACY TEST
ASSUMPTIONS CHECK
AQI MAP USING TABEAU PUBLIC

Figure 2: Flow Chart Model

The general cycle comprises of various advances:

1) Data assortment: There is a different method from which we collected data from various dependable sources like Delhi Gov. site.

2) Exploratory examination: We research and explore examination with various parameter like ID of outliners, consistency check, missing qualities, and so on, it's totally occurred in this period of the venture.

3) Data Manipulation control: In period of data control stage the required missing data need to insert in utilizing the mean estimations of that characteristic of information. [2]

4) Prediction of boundaries utilizing by gauge model: For appropriate data indirect relapse we have to keep future qualities for different boundaries just Example, the degree of nitrogen oxide (NO2), NH2 etc.

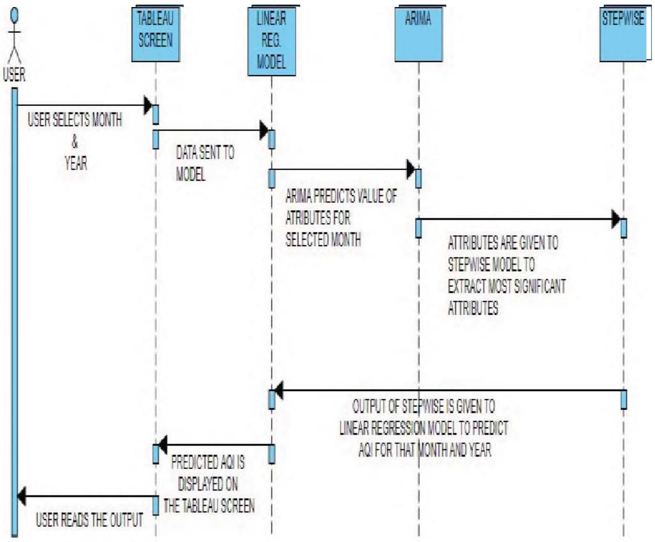
5) Implementation of straight relapse: Whenever all the boundaries become in active mode or they are accessible mode, the direct relapse calculation would be used in anticipate the air quality index (AQI).

6) Data accuracy investigation: We have to analyze that used model is being fit for overall data or not so we have to cross check root mean error, absolute percentage error then after we have to assume this factor is good for accuracy or not

C. Sequence Diagram:

When the end user enters the either particular month, year or date. the session requests us generated and information sent to back.

The ARIMA predicts various boundaries for that month, during this stepwise data discovering foreseeing the AQI. And AQI send back information and shown back to the client screen.


USER
TABLEAU SCREEN
LINEAR REG. MODEL
ARIMA
STEPWISE
USER SELECTS MONTH & YEAR
DATA SENT TO MODEL
ARIMA PREDICTS VALUE OF ATRIBUTES FOR SELECTED MONTH
ATTRIBUTES ARE GIVEN TO STEPWISE MODEL TO EXTRACT MOST SIGNIFICANT ATTRIBUTES
OUTPUT OF STEPWISE IS GIVEN TO LINEAR REGRESSION MODEL TO PREDICT AQI FOR THAT MONTH AND YEAR
PREDICTED AQI IS DISPLAYED ON THE TABLEAU SCREEN
USER READS THE OUTPUT

IV. IMPLEMENTATION

A. Dataset

Dataset/Source: Delhi Govt. Website

Structured/Unstructured data: I have taken Structured Data in CSV format.

Dataset Description:

Pollutants data is collected from Delhi state.

http://aqicn.org/city/delhi

So we had 20383 records. This dataset consist of 13 data values listed below.

1) *station_code*
2) *sampling_data_date*
3) *state_*
4) *location_*
5) *agency_name*
6) *type_R/I*

7) *so2(sulfer_dioxide)*
8) *no2_*
9) *rspm_*
10) *spm _*
11) *location_monitoring_station_*
12) *pm2_5*
13) *date_*

Information fields and their depiction: -

1. The Air Quality Index (AQI) [9]: is define about an inventory which contain record of the ordinary air greatness. Polluted and clean air data arises here every day by some parameter.

2. STATION: station define where we have enrolled vehicle in different place in Delhi.

3. Time & DATE: The gathered information incorporates month to month normal information for various areas of New Delhi for a very long time.

4. ZONE: We have collected the data of stations in various zone to be specific in city which we consider as East, West, North & south Delhi.

B. *Pollution Components*

- PM10: If the numerous particles come in the range of 2.5 and 10 micrometers then it's come under category of PM10 and it's in form of dust, spores, seethe etc.
- PM2.5: PM 2.5 components comes in range of less than 2.5 micrometers or less in distance across. This particle produces from like car vehicles, power plants, uptown wood consuming, backwoods fires, so on.•
- CO: CO Gas produces when incomplete burning of carbon-containing fuels, such as coal, oil, charcoal, wood, kerosene, natural gas and propane. exhaust, fuel combustion etc.
- NH3: NH3 is present approx. 70 % in environment.
- NO2: Nitrogen Dioxide (NO2) is a one of very refluent gases normally known as Nitrogen oxide, it's produces in air from burning of fuel, emissions from vehicles like car, tractor, trucks, bigger power plants and other road equipment.

V. Data Chunking

A. *There is a very important step in project to chunking of data, There is a very important role of data control In machine learning model*

1) Collecting of Station Data: We have taken 10 recent station data of various geographical in this algo.

2) Linear Regression Model: Linear Regression is a sciectical approach of between dependent variables and independent varibles which shows a modeling relationship between both.a ration of 80:20.

3) Basically step wise regression is taken for various variable selection to judgments the significant variable attribute which is used in this model. [3]

4) Stepwise Multiple Linear Regression Method:
Each Data science or engineer start out with initial linear regression as a fist algorithm, because it is basic method and very important thing to tackle complex machine learning problems. In multiple linear regression produces one output variable and multiple input variables.

The output result came from combination of attribute parameter like PM10, NH3, NO2 and of course no of registered vehicle in various station. On checked the blend of factors returned by step astute for P esteems the p esteems are under 0.05 Therefore, we will utilize can utilize this model for foreseeing AQI. [7]

B. *Checking forAssumptions:At last we should check if the presumptions ofstraight*

1) Normality of Residuals: At last we need to check if the assumption of linear regression is meet or not if meet then how much percentage. First We need to check normality of residual, Actual we shape a curve in behalf of plotting the residual curve value and training value.

2) Other assumption we can calculate from multi collinearity, in this concept independent variables should not establish mutually connection between each other.

3) linearity check: We can identify linearity when we draw the residual values against of various independent attributes variables. Thus assumption can be checked with a histogram or a Q-Q-Plot.

VI. Python and Jupyter Tool :

We have been used python and Jupiter tool in this project

1) 1) Python: Python is an interpreter, high programming language. It's an open access Python IE, With the help of this software we can use for writing source code for front end and it's send a request to back end and back end get a request from from end as accordingly redirect response to inter mediate layer.

2) Jupyter Notebook: The Jupyter Notebook is a open aces free web application tool that give you facilitates to make, edit and share the records that contain live code, examinations, representations and spellbinding data. Uses include: information cleaning and change, mathematical reproduction, factual demonstrating, information representation, AI, and some more. [4]

VII. RESULTS

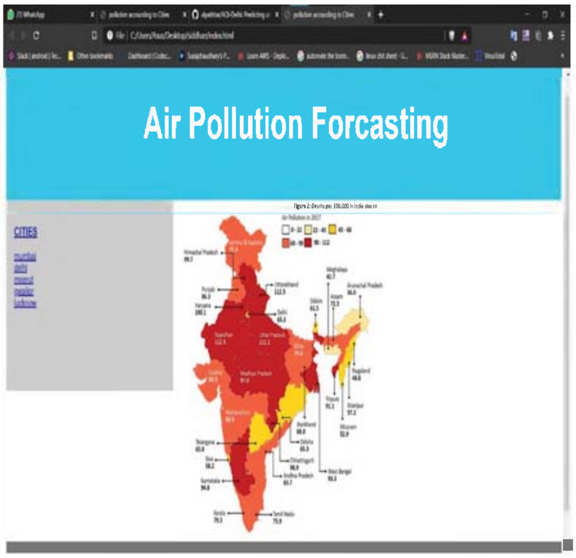
We have taken four assessors to evaluates correct effectiveness execution in this model. They are:

- ROOT MEAN SQUARE ERROR (RMSE):
- MEAN ABSOLUTE ERROR PERCENTAGE: The
- MAPE
- MEAN ABSOLUTE ERROR:

The mean outright mistake (MAE) is basic relapse blunder metric to grasp.

We will figure the remaining for each information point, taking just the total estimation of each, so that negative and positive residuals don't counterbalance.

VIII. ANALYSIS



On continuing with our Air quality prediction project we recently making this first page description how we mapped the cities on the left (◀) side by clicking on these cities we get the AQI of particular city and after getting the AQI of particular city it's shows its harms for our health or not and then we decided to survive or not in this city properly.

IX. SOURCE CODE

```
* {
  box-sizing: border-box;
}

body {
  font-family: Arial, Helvetica, sans-serif;
}

/* Style the header */
header {
  background-color: rgb(17, 204, 236);
  padding: 30px;
  text-align: center;
  font-size: 35px;
  color: white;
}

/* Create two columns/boxes that floats next to each other */
nav {
  float: left;
  width: 30%;
height: 300px; /* only for demonstration, should be removed */
  background: #ccc;
  padding: 20px;
}

/* Style the list inside the menu */
nav ul {
  list-style-type: none;
  padding: 0;
}

/* Clear floats after the columns */
section:after {
  content: "";
  display: table;
  clear: both;
}

/* Style the footer */
footer {
  background-color: #777;
  padding: 10px;
  text-align: center;
  color: white;
}

/* Responsive layout - makes the two columns/boxes stack
on top of each other instead of next to each other
on small screens */
@media (max-width: 600px) {
  nav,
  article {
     width: 100%;
     height: auto;
  }
}
```

X. CONCLUSION

Precision of our model is very acceptable. The anticipated AQI has a precision of 96%. Future upgrades incorporate expanding the extent of district and to incorporate whatever number locales as could be allowed as of now this venture targets foreseeing the AQI estimations of various areas of close by New Delhi. Further, by utilizing information of various urban areas the extent of this venture can be exhausted to anticipate AQI for different urban communities also.

References

[1] Ni, X.Y.; Huang, H.; Du, W.P. "Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data." Atmos. Environ. 2017, 150, 146–161.

[2] G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," Environ. Model. Softw., vol. 80, pp. 259-264,2016.

[3] 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).

[4] Suhasini V. Kottur , Dr. S. S. Mantha. "An Integrated Model Using Artificial Neural Network

[5] 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

[6] 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

[7] 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

[8] PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG," Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations," IEEE ACCESSJuly 30, 2018.Digital Object Identifier10.1109/ACCESS.2018.2849820.

[9] GaganjotKaur Kang, Jerry Zeyu Gao, 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, January2018.

[10] Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner and Daniel Price, "Air Quality Simulations using Big Data Programming Models," IEEE Second International Conference on Big Data Computing Serviceand Applications,2016.