Pollution Parameter Analysis and Predictive Model Building

Roopa G K[1] Bharathi K[2]
[1,2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]VCET, Puttur, India

Abstract— Monitoring and preserving air quality has become one of the most essential activities in many industrial and urban areas today. The quality of air is adversely affected due to various forms of pollution caused by transportation, industries, fuel uses etc. The deposition of harmful gases is creating a serious threat for the quality of life in smart cities. With increasing air pollution, we need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. Hence, air quality evaluation and prediction has become an important research area. The quality of air is affected by multi-dimensional factors including location, time, and uncertain variables. Recently, many researchers began to use the big data analytics approach due to advancements in big data applications and availability of environmental sensing networks and sensor data. The aim of our project is to implement various big-data and machine learning based techniques for air quality forecasting. For this, the front end is designed using HTML and CSS, back end is designed using PHP and MySQL is used for the data storage. The algorithm implementation and Predictive Model building is done using Python.

Keywords: Predictive Model Building, HTML, CSS, PHP, MySQL, Python

I. INTRODUCTION

Air quality being a concern for all, the knowledge on instantaneous air quality data has become a need for the society. The instantaneous data may also become relevant for depicting real-time generic information in the form of Air Quality Index(AQI). This has been taken as one of the priority of concern for Central Pollution Control Board(CPCB). These data once started displaying at respective locations, would bring participation of general public in the prevention & control of air pollution at respective area of activity. The data may also link with all developmental authorities in respective areas.

Over the years, CPCB have established a large no. of on-line real time Continuous Ambient Air Quality Monitoring Stations(CAAQMS) across the country and gradually these data are being taken for use by the regulatory organizations and attempts are on to link these data with developmental authorities / local bodies / road transport authorities. Thus, stress is required for checking of all data submitted / generated by concerned agencies, so that the generated data may be linked with developmental authorities at local areas without any question mark. These data may also be used for regional planning. The CPCB has prepared National Air Quality Index, wherein all real-time air quality data will be used for projection of real-time AQI at all respective locations. Thus, the data generated by CAAQMS have to be made available at respective sites with scientific validation so that the same may be utilized with confidence.

It's true that viewing and validating all these data for all stations is not a feasible option. Hence, we need to develop a system with in-built features so that all (regulators, local authorities, public in general etc.) can utilize the data at respective levels besides review of actions plan for all the area with respect to air quality goal. The in-built system should be such that all will love to work with maximum efficiency in minimum time on regular basis.

It may be remembered that instantaneous data dissemination is not considered as validated data dissemination to the public. The actual validation of data is done by a team of experts in a regular frequency. While doing the validation process, the experts discuss the issues related to local conditions, meteorological data, rationality on data of the past, exceedance in data range, calibration frequency, outlier data etc. On satisfactory validation of data, the validated data is published by the experts are acceptable for all. To maintain transparency of the system, collaborative portal for handling of issues are though mandatory.

II. EXISTING SYSTEM

A commonly used numerical method is the Commonly Multi-Scale Air Quality(CMAQ) modeling system. Since the 1970s,three generations of CMAQ models have been developed. Lou et al used the CMAQ modeling system to analyze and evaluate air pollutant ozone concentrations in China and proposed that this method could be applied to other oxides of Nitrogen.

Potential forecasting was widely used in the 1960s. It is based on combination of statistical and numerical prediction, according to certain conditions of the pollution source, the weather situation, and the meteorological index to construct the pollution potential index (PPI), and the qualitative or semi quantitative forecast of the atmospheric environment quality in the future is conducted. When the weather conditions are expected to be in line with the standards of possible serious pollution, a warning is issued. Scott et al. applied an air pollution potential forecast model based on a synoptic climatological approach to forecast the concentration of SO2 in a heavily industrialized area in Durban (South Africa). Their proposed model identified periods of elevated SO2 successfully. However, potential prediction without considering the location of pollution sources and emissions of pollutants and the accuracy of prediction is low. Therefore, more statistical models, artificial intelligence, and hybrid models are used currently.

Statistical forecast methods analyse the events without knowing the mechanism of the change; therefore, this method is not dependent on physical, chemical, or biological processes. Statistical forecasts include parametric and non-parametric statistical methods. Parametric models are traditional statistical models such as linear regression and principal component analysis; nonparametric models have no defined function form. Generally, nonparametric regression

includes neural network models, Gaussian process regression etc.

III. PROPOSED SYSTEM

In the proposed method three types of data sets are being collected:

- Air Pollution data from The Central Pollution Control Board.
- Landmark information around Sensors through Google APIs.
- Health Statistical data from the Regional Health Centres.

A descriptive analysis is performed on these data's. A prediction model is built and tested for each city on various pollution parameters. Then the effect of the variations in these parameters is correlated with the health statistics of the people in those regions.

A. Phase 1: Predictive Model Building.

1) Stage 1: Data Extraction

Data is extracted from The Central Pollution Control Board's website through web scrapping in Python using Selenium and BeautifulSoup module. The extracted data is exported as a Comma Separated File in to excel sheet.

2) Stage 2: Data Cleaning

The extracted data is cleaned using pandas. DataFrame.dropna( ) library in python.

3) Stage 3: Descriptive analysis

We are performing a Descriptive analysis on the cleansed data using the describe( ) function in python that calculates the Mean, Standard Deviation, Mode and also we are using corr( ) and skew( ) functions to calculate the correlation and skewness of data respectively.

4) Stage 4: Data Modelling

Using the Random Forest Classification Method we are building a prediction model, which can predict the value of meteorological parameters from the historical values.

5) Stage 5: Estimation of Performance

The performance of the model is evaluated using the Gini coefficient, Cross Validation and Root Mean Square methods.

B. Phase 2: Location Information Collection

Information regarding the minimum, maximum and average distances of Schools, Colleges, Public Bus Stands, Railway Stations and Industries from the sensors is collected through Google APIs and the impact of various air pollutants is analysed.

C. Phase 3: Webpage development

A webpage is developed using HTML and CSS as the front end design and PHP for the back end. On entering the city name the user will get information about the concentration of various air pollutants in that region, in a user friendly way.

D. Phase 4: Disease Analysis

The correlation between variations in air pollutants and health statistics of the native people is analysed and the results are displayed on the webpage through user interfaces.

IV. WORKFLOW OF SYSTEM

Fig. 1: Complete System Workflow Diagram

V. ALGORITHM FOR LOGIC IMPLEMENTATION

We are using PCA algorithm in our project. The steps involved in PCA algorithm are:

1) Standardize the data
2) Calculate the covariance matrix
3) Compute the Eigen vectors and Eigen values of the covariance matrix to identify the principal components.
4) Feature vector

VI. SCREENSHOTS

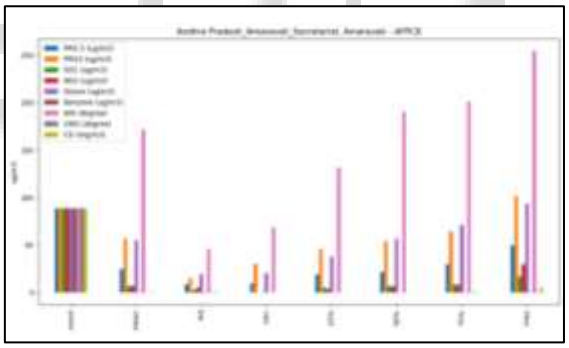

Fig. 2: Description of Data

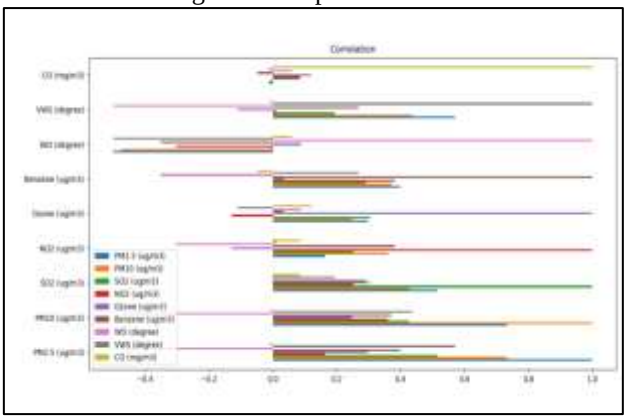

Fig. 3: Correlation between the data

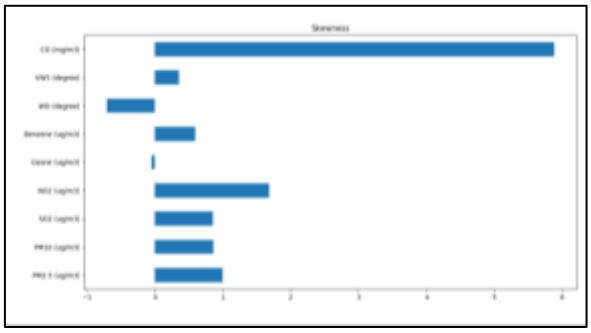

Fig. 4: Skewness of each attribute

REFERENCE

[1] Alka Yadav and DurgaToshniwal- "Extracting Patterns and Variations in Air Quality of Four Tier I Cities in India".2017 IEEE 13th World Congress on Services.978-1-5386-2002-1/17$31.00copyright2017 IEEE 17-20.

[2] Yash Mehta, ManoharaPai M.M., SanoopMallissery and Shwetanshu Singh- "Cloud enabled Air Quality Detection, Analysis and Prediction - A Smart City Application for Smart Health".2016 3rd MEC International Conference on Big Data and Smart City.978-1-4673-9584-7/16/$31.00 copyright2016 IEEE.

[3] Zheng Cheng-Jian, Li Xian-Hua, Huang-Wei, Shi Xiang-Yong, Luo Qing-Zhou- "Analysis on the Correlation of Atmospheric Path Radiation and Air Pollution Index".2009 Urban Remote Sensing Joint Event.978-1-4244-3461-9/09/$25.00 ©2009 IEEE

[4] P Geetha and M Kokila- "Estimation of Air Pollution using Remote Sensing Technique in Coimbatore-A Case Study".IEEE ICCSP 2015 Conference.978-1-4799-8081-9/15/$31.00 copyright2015 IEEE 0794-0798.

[5] YulianYangl, Kun Yangl, YanhuiZhul, Yi He-"The Seasonal Characteristics and Spatial Distribution of PM2.5 in China". Supported by National 863 Program (No. 2012AA121402), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20115303110002), the Research of Yunnan Province Social Science Fund Projects (No.2010CA010).

