International Journal of Engineering Innovations in Advanced Technology
ISSN: 2582-1431 (Online), Volume-4 Issue-2, June 2022

Air Quality Prediction Using Machine Learning

K Jayith Reddy, Y Subodh
UG Students, Department Of CSE, Sphoorthy Engineering College, Hyderabad, India.
Email: k.jayithreddy123@gmail.com, subodh9100@gmail.com

Mohd Afzal
Assistant Professor, Department Of CSE, Sphoorthy Engineering College, Hyderabad, India.
Email: mdafzal.mbnr@gmail.com

Abstract **- Air quality of a certain region can be used as one of the major factor determining pollution index also how well the city's industries and population is managed. Urban air quality monitoring has been a constant challenge with the advent of industrialization. Air pollution has remained a major challenge for the public and the government all over the world. Air pollution causes noticeable damage to the environment as well as to human health resulting into acid rain, global warming, heart diseases and skin cancer to the people. This project addresses the challenge of predicting the air quality index (AQI), with the aim to minimize the pollution before it gets adverse, using two machine learning algorithms: neural networks and support vector machines. The air pollution databases were extracted from the central pollution control board (CPCB), ministry of environment, forest and climate change, government of India. The proposed machine learning (ML) model is promising in prediction context for the Delhi AQI. The results show improvement of the prediction accuracy and suggest that the model can be used in other smart cities as well.**

Keywords **- Crop Diseases, Agriculture, Artificial Intelligence, Cloud, CNN, Mobile, Plant Pathology, Neural Networks**

I INTRODUCTION

The Environment is nothing but everything that encircles us. The environment is getting polluted due to human activities and natural disaster, very severe among them is air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. If the humidity is more, we feel much hotter because sweat will in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO2 etc. Carbon Monoxide is produced due to the deficient Oxidization of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel; Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems; Nitrogen oxides causes dizziness, nausea; Particulate matter with a diameter 2.5 micrometer or less than that affects more to human health. Measures must be taken to minimize air pollution in the environment. Air Quality Index (AQI), is used to measure the quality of air. Earlier classical methods such as probability, statistics were used to predict the quality of air, but those methods are very complex to predict the quality of air. Due to advancement of technology, now it is very easy to fetch the data about the pollutants of air using sensors. Assessment of raw data to detect the pollutants needs vigorous analysis. Convolution Neural networks, Recursive Neural networks, Deep Learning, Machine learning algorithms assures in accomplishing the prediction of future AQI so that measures can be taken appropriately. Machine learning which comes under artificial intelligence has three kinds of learning algorithms, they are the Supervised Learning, Unsupervised learning, Reinforcement learning. In the proposed work we have used supervised learning approach. There are many algorithms under supervised learning algorithms such as Linear Regression, Nearest Neighbor, SVM, kernel SVM, Naive Bayes and Random Forest. Compared to all other algorithms Random Forest gives better results, so our approach selects Random Forest to predict the accurate air pollution

II LITERATURE REVIEW

Ishan et.al [1] described the benefits of the Bidirectional Long - Short Memory [BiLSTM] method to forecast the severity of air pollution. The proposed technique achieved better prediction which models the long term, short term, and critical consequence of PM2.5 severity levels. In the proposed method prediction is made at 6h, 12h, 24h. The results obtained for 12h is consistent, but the result obtained for 6h, and 24h are not consistent. Chao Zhang et.al [2] proposed web service methodology to predict air quality. They provided service to the mobile device, the user to send photos of air pollution. The proposed method includes 2 modules a) GPS location data to retrieve the assessment of the quality of the air from nearby air quality stations. b) they have applied dictionary learning and

Manuscript received June 15, 2022; Revised June 25, 2022; Accepted June 30, 2022

convolution neural network on the photos uploaded by the user to predict the air quality. The proposed methodology has less error rate compared to other algorithms such as PAPLE, DL, PCALL but this method has a disadvantage in learning stability due to this the results are less accurate.

Ruijun Yang et.al [3] used the Bias network to find out the air quality and formed DAG from the data set of the town called as shanghai. The dataset is dived for the training and testing model. The disadvantage of this approach is they have not considered geographical and social environment characteristics, so the results may vary based on these factors. TemeseganWalelignAyeleet.al [4] proposed an IoT based technique to obtain air quality data set. They have used Long Short-term Memory [LSTM] technique in- order to predict the air quality the proposed technique achieved better accuracy by reducing the time taken to train the model. But still, the accuracy can be improved by compared other techniques such as the Random forest method NadjetDjebbriet.al [5] proposed artificial based Regressive model which is nonlinear to predict 2 major air pollutants such as carbon monoxide and nitrogen oxides. They have considered the variables such as the speed of the air, air direction, temperature, and moisture and the toxic elements from the industrial site such as Skikda. They have used RMSE and MAE to determine the performance, but this method considered only 2 pollutants such as NO and CO the other major pollutants such as sulfur dioxide, PM2.5, PM10 are not considered.

III SYSTEM DESIGN

A SDLC Models

There are various software development life cycle models defined and designed which are followed during the software development process. These models are also referred as Software Development Process Models". Each process model follows a Series of steps unique to its type to ensure success in the process of software development.

Following are the most important and popular SDLC models followed in the industry –

- Waterfall Model
- Iterative Model
- Spiral Model
- V-Model
- Big Bang Model

Other related methodologies are Agile Model, RAD Model, Rapid Application Development and Prototyping Models. The spiral model combines the idea of iterative development with the systematic, controlled aspects of the waterfall model. This Spiral model is a combination of iterative development process model and sequential linear development model i.e. the waterfall model with a very high emphasis on risk analysis. It allows incremental releases of the product or incremental refinement through each iteration around the spiral.

B Spiral Model – Design

The spiral model has four phases. A software project repeatedly passes through these phases in iterations called Spirals.

C Identification

This phase starts with gathering the business requirements in the baseline spiral. In the subsequent spirals as the product matures, identification of system requirements, subsystem requirements and unit requirements are all done in this phase. This phase also includes understanding the system requirements by continuous communication between the customer and the system analyst. At the end of the spiral, the product is deployed in the identified market.

D Design

The Design phase starts with the conceptual design in the baseline spiral and involves architectural design, logical design of modules, physical product design and the final design in the subsequent spirals.

E Construct or Build

The Construct phase refers to production of the actual software product at every spiral. In the baseline spiral, when the product is just thought of and the design is being developed a POC (Proof of Concept) is developed in this phase to get customer feedback. Then in the subsequent spirals with higher clarity on requirements and design details a working model of the software called build is produced with a version number. These builds are sent to the customer for feedback.

F Evaluation and Risk Analysis

Risk Analysis includes identifying, estimating and monitoring the technical feasibility and management risks, such as schedule slippage and cost overrun. After testing the build, at the end of first iteration, the customer evaluates the software and provides feedback.

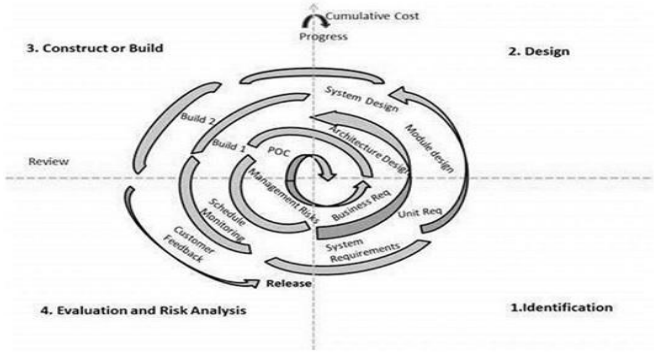


Figure 1 SLDC

The following illustration is a representation of the Spiral Model, listing the activities in each phase.

Based on the customer evaluation, the software development process enters the next iteration and subsequently follows the linear approach to implement the feedback suggested by the customer. The process of iterations along the spiral continues throughout the life of the software.

F Existing System

In first approach monitoring of real-time Air Quality Monitoring and another is developing statistical models using historical data. This project summarizes Air Quality Prediction studies in two categories. The first study is on prediction of PM2.5 / PM10 concentration and on prediction of air pollutants like CO2, O3,NO2 and then inferring Air Quality Index (AQI) using machine learning techniques whereas second study is on monitoring real-time AQI using sensor devices.

Disadvantages

We consider only some of the parameters. Here we can only find the statistical values

G Proposed System

The benefit of linear regression is that it is the simplest repressor and is readily interpretable. Linear Regression 20 is used for finding linear relationship between target and one or more predictors. It is generally applied to one independent variable and one dependent variable. To train a model we first check how well the model fits the training data. Its parameters are then set to fir the training set. Generally, RMSE value is used to evaluate the performance of a regression model P9P652del.

Advantages

We consider all the parameters. Here we can find the accurate values.

IV SYSTEM DESIGN

A Data Flow

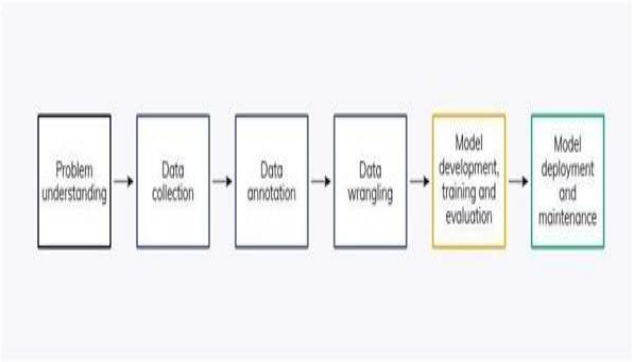


Figure 2 Data Flow

B System Architecture

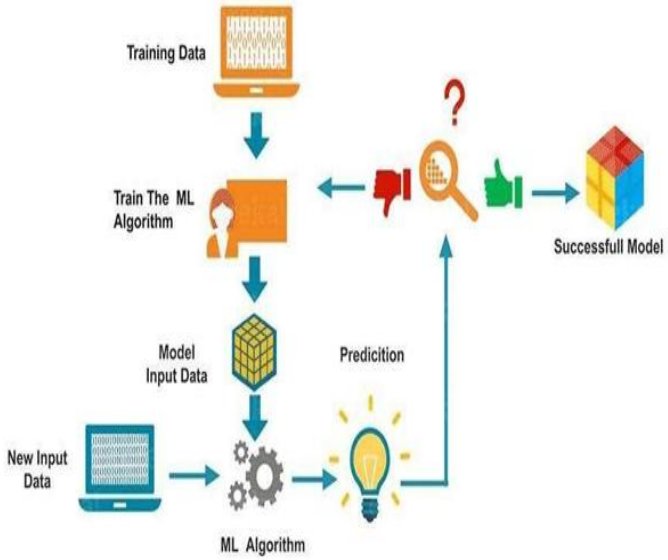


Figure 3 Architecture

C Flow Architecture

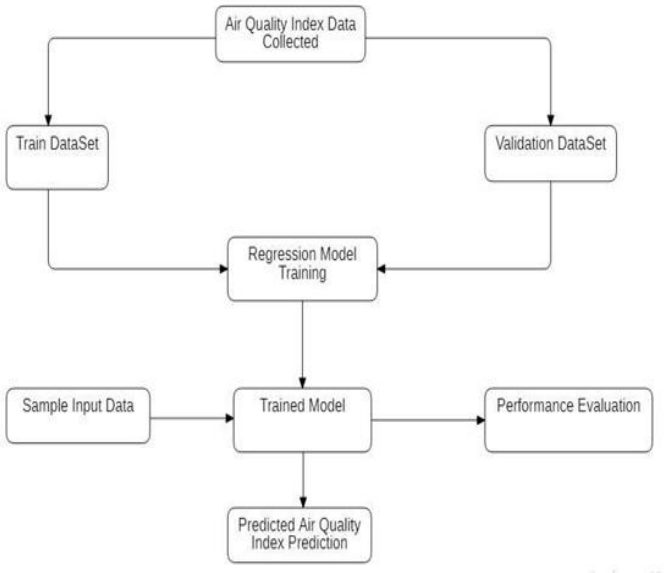


Figure 4 Flow Architecture

D Input Design and Output Design

Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the

data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

What data should be given as input?

How the data should be arranged or coded?

The dialog to guide the operating personnel in providing input.

Objectives

1. Input Design is the process of converting a user-oriented description of the input into a computer- based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively.
2. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

3. Select methods for presenting information.

4. Create document, report, or other formats that contain information produced by the system.

5. The output form of an information system should accomplish one or more of the following objectives. Convey information about past activities, current status or projections of the Future.

6. Signal important events, opportunities, problems, or warnings.

7. Trigger an action.

8. Confirm an action.

E Data Pre-Processing

The final data fed to the model is created by merging two datasets (pollution.csv and weather.csv) based on time. From observing the datasets, it is found that the two datasets have different number of observations per hour.

The weather data has 34446 records observed every 20 mins from 2017-01-01 00:20:00 to 2019-01-01 23:50:00.

The pollution data has 17520 observations per species observed every hour from 2017-01-01 00:00:00 to 2018-12-31 23:00:00 for 3 different species (NO2, O3, PM2.5).

Step 1: Parsing Date Time

Since the data is read from a csv file, the data type of Reading Date Time was not in DateTime format, hence it was parsed into DateTime type while reading the csv. The best approach would be to store the data sets as a pickle file(.pkl) which would preserve the data type of the features.

```
#Parsing date column of both datasets to be read by python

pdata = pd.read_csv('data/pollution-1.csv',parse_dates=['ReadingDateTime'])
wdata = pd.read_csv('data/weather-1.csv',parse_dates=['DATE'])
```

Step 2: Extracting numerical values

```
#Extracting numerical values of dew point, temperature, wind speed (converting from mps to kmph)
#and direction

wdata['DEW'] = wdata['DEW'].str[:-2].astype(np.float64)/10
wdata['TMP'] = wdata['TMP'].str[:-2].astype(np.float64)/10
wdata['DIR'] = wdata['WND'].str[:3].astype(np.float64)
wdata['SPD'] = (wdata['WND'].str[8:-2].astype(np.float64)/10)*3.6

#Calculating relative humidity from dew point and temperature

wdata['HUM'] = 100*(np.exp((17.625 * wdata['DEW'])/(243.04 +
wdata['DEW']))/np.exp((17.625 * wdata['TMP'])/(243.04 + wdata['TMP'])))

#Replacing missing wind direction observations with nulls

wdata.DIR.replace(999,np.nan,inplace=True)
```

The features in the weather dataset are not explicitly numerical values; each value is a reference to a set of information through which the actual numerical values are to be extracted. This information is detailed in the following code block and table 2

Table 2 Numerical Value Extraction

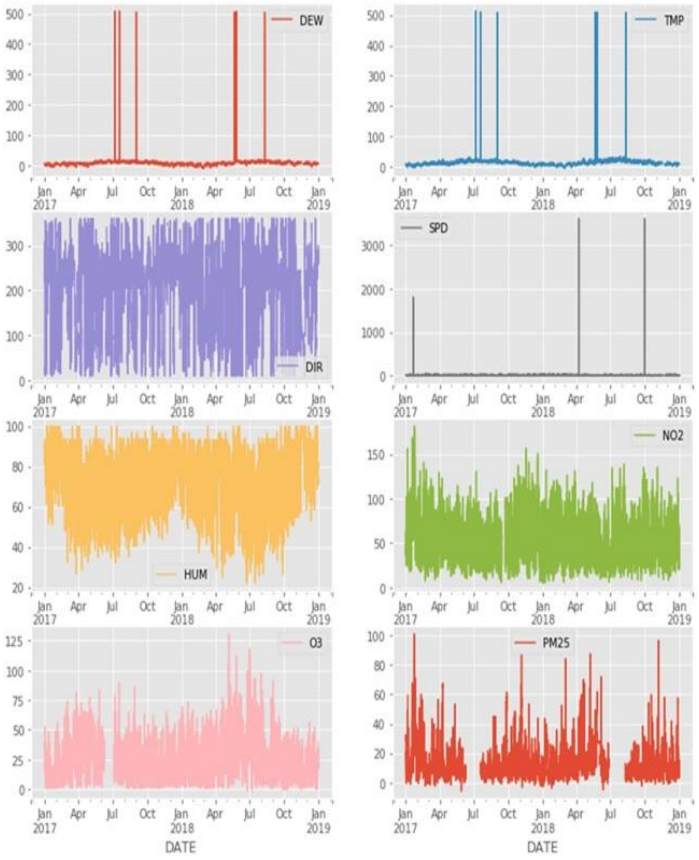
Name	String Index	Description
DEW	[0-4]	The observed dew point temperature value with the following properties **Min:** -0982 **Max:** +0368 **Units:** Degrees Celsius **Scaling Factor:** 10
DEW	[5-6]	The code that denotes a quality status of the reported dew point temperature.
TMP	[0-4]	The observed temperature value with the following properties: **Min:** -0932 **Max:** +0618 **Units:** Degrees Celsius **Scaling Factor:** 10
TMP	[5-6]	The code that denotes a quality status of the reported temperature.
WIND	[0-2]	The angle, measured in a clockwise direction, between true north and the direction from which the wind is blowing. **Min:** 001 **Max:** 360 **Units:** Angular Degrees **Scaling Factor:** 1 **Missing:** 999
WIND	[4]	The code that denotes the quality status of a reported wind direction angle.
WIND	[6]	The code that denotes the character of the wind observation, which is a qualitative metric to indicate calm or strong winds.
WIND	[8-11]	The observed wind speed which is the rate of horizontal travel of air past a fixed point. **Min:** 0000 **Max:** 0900 **Units:** meters per second **Scaling Factor:** 10
WIND	[13]	The code that denotes the quality status of the reported wind speed.

Relative Humidity: It the ratio between the amount of water vapour at a given temperature to the maximum amount of water vapour the air can hold at that temperature. This is calculated from DEW point and TMP values.

Wind Values: The wind speed is converted from metres per second to kilometres per hour. The reference for the weather data set also informs that wind direction= 999 indicates a missing value. These are replaced to be nulls.

IV DATA VISUALIZATION

The time series visualization of the final data set was plotted to construct initial impressions about the collected data.



CONCLUSION

Throughout this project, several models which can predict Pm 2.5 levels and classify them into different pollution bands were experimented and their performance was successfully evaluated. The exploratory data analysis and feature engineering methods implemented for the prediction models revealed interesting correlations between weather and pollution data. We obtained several notable outcomes from the predictive mode ls that are worth being discussed. Different approaches to handle null values yielded varied performance from each of the models, however simply dropping the records that had null values seemed to be the best approach. Between obtaining the AQI by predicting the PM2.5 values and using a classifier to predict the AQI band straight away, the classifier seemed to perform better. A regression model could be used for applications in data analytics, but it is concluded that classifier models perform better for air quality prediction.

REFERENCES

1. N. McCrea, "An Introduction to Machine Learning Theory and Its Applications", A Visual Tutorial with Examples.
2. https://www.forbes.com/sites/davidteich/2018/12/26/machine-learning-and-artificial- intelligence-in-business-year-in-review-2018/#980755b2041c.
3. P. Kavitha and M. Usha, "Anomaly Based Intrusion Detection In WLAN Using Discrimination Algorithm Combined with Naive Bayesian Classifier & quot;, Journal of Theoretical and Applied Information Technology, vol. 62, no. 3, pp. 646-653, 2014.
4. T. Ensari, M. Günay, Y. Nalçakan and E. Yildiz, & quot; Overview of Machine Learning Approaches for Wireless Communication & quot;, InNext-Generation Wireless Networks Meet Advanced Machine Learning Applications, pp. 123-140, 2019.
5. J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, et al., "Feature selection: A data perspective", ACM Computing Surveys (CSUR), vol. 50, no. 6, pp. 94, Jan 2018.
6. Ledesma, G. Cerda, G. Aviña, D. Hernández and M. Torres, "Feature selection using artificial neural networks", InMexican International Conference on Artificial Intelligence, pp. 351-359, 2008 Oct 27.
7. B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines", European Journal of Operational Research, vol. 265, no. 3, pp. 993-1004,Mar 2018.
8. M. Shahbaz, S. A. Taqvi, A. C. Loy, A. Inayat, F. Uddin, A. Bokhari, et al., "Artificial neural network approach for the steam gasification of palm oil waste using bottom ash.