Indian Air Quality Analysis & Prediction

B.TECH SEM – VII Mini PROJECT
Dept. of Computer Science & Engineering

By

Umang Mehta 18BCP120
Sahil Monpara 18BCP095
Dhara Barot 18BCP142D

<u>Under the Supervision</u>
<u>Of</u>
Dr. Nayantara Kotoky

SCHOOL OF TECHNOLOGY
PANDIT DEENDAYAL ENERGY UNIVERSITY
GANDHINAGAR, GUJARAT, INDIA

July – December, 2021

Abstract

Predicting and presenting air quality is necessary step to be taken by government as it is becoming the major concern among the health of human beings. Various air pollutants causing air pollution are Carbon dioxide, Nitrogen dioxide, carbon monoxide etc that are released from burning of natural gas, coal and wood, industries, vehicles etc. Air Pollution can cause severe disease like lungs cancer, brain disease and even lead to death. Air quality Index measure the quality of air. Machine learning algorithms helps in determining the air quality index. Various research is being done in this field but still results are not accurate and good prediction web app is not present in public domain. Dataset are available from Kaggle, air quality monitoring sites and divided into two Training and Testing. Machine Learning algorithms employed for this are Linear Regression, Decision Tree, Random Forest, Artificial Neural Network, Support Vector Machine. Among all, decision tree and random forest are useful for further research in our project. We make web app to describe the prediction data and practical application of that data.

Keywords: Air quality Index, Decision Tree Regression, Random Forest Regression, Flask, Data Visualization

Table of Contents

Chapter - 1: Introduction

Air Pollution occurs when harmful or excessive quantities of substances including gases, particles, and biological molecules are introduced into the Earth's atmosphere. Air pollution in India is a serious issue, ranking higher than smoking, high blood pressure, child and maternal malnutrition, and risk factors for diabetes. At least 140 million people breathe air 10 times or more over the WHO safe limit and 13 of the world's 20 cities with the highest annual levels of air pollution are in India. Air pollution contributes to the premature deaths of 2 million Indians every year. In urban areas, most emissions come from vehicles and industry, whereas in rural areas, much of the pollution stems from biomass burning for cooking and keeping warm. In autumn and winter months, large scale crop residue burning in agriculture fields – a low cost alternative to mechanical tilling – is a major source of smoke, smog and particulate pollution. Worsening air pollution has been amongst India’s most pressing problems in recent years. In 2019, air pollution led to about 12.5 percent of all deaths in the country. In the same year, it resulted in an economic loss of approximately 1.4 percent of GDP. According to IQAir, in 2020, India ranked third amongst all countries in the world with the worst air quality. Data analysis and prediction can create awareness in the different section of the society.

We want to create a platform like a Web-App using Flask where people can understand how these features are related to AQI(Air Quality Index). And also predict the AQI using values of these features. We also want to visualize the AQI values according to states, cities, and areas.

Chapter - 2: Literature Review

1) Indian Air Quality Prediction and Analysis using Machine Learning

Air quality index of India is a standard measure used to indicate the pollutant (so2, no2, rspm, spm. etc.). air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient descent boosted multivariable regression problem. 95% accuracy in this particular model. 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) **[1].**

2) Air Quality Prediction using Machine Learning Algorithm

Two types of Pollutants that are causing air pollution are Primary Pollutants and Secondary Pollutants.
Primary Pollutants : Carbon dioxide (CO2),Sulphur oxide (SOX),Nitrogen oxide (NOX), Carbon monoxide (CO),
Toxic metals : Lead and Mercury
Secondary Pollutants : Ground Level Ozone , Acid Rain
Primary Pollutants are those which are released into air directly from Source whereas Secondary Pollutants are those which are formed by reacting with either primary pollutants or with other atmospheric components.
Various algorithms had taken meteorological data like temperature, wind speed, humidity in predicting accurately the upcoming pollutant level. Neural Network and boosting model comes out to be superior than other algorithms **[2]**.

3) Data Mining to Aid Policy Making in Air Pollution Management

Taiwanese government has set up the National Air Quality Monitoring Network (TAQMN) to monitor nationwide air quality and adopted an array of measures to combat this problem.
The mining process consists of data acquisition from Web sites of 71 data gathering stations nationwide, data pre-processing using multi-scale wavelet transforms, data pattern identification using cluster analysis, and final analysis in mapping the identified clusters to geographical locations **[3]**.

4) Review of Air Quality Monitoring: Case Study of India

Identify critical problem areas suffering from severe air pollution by an objective assessment of state of practice and to recommend suitable measures for improvement wherever applicable.
In this paper ten major cities have been taken into consideration for air quality monitoring. These cities have been chosen bearing in mind the fact that these cities are major centres for commercial, industrial and tourist activities due to which the rate of environmental deterioration is high.

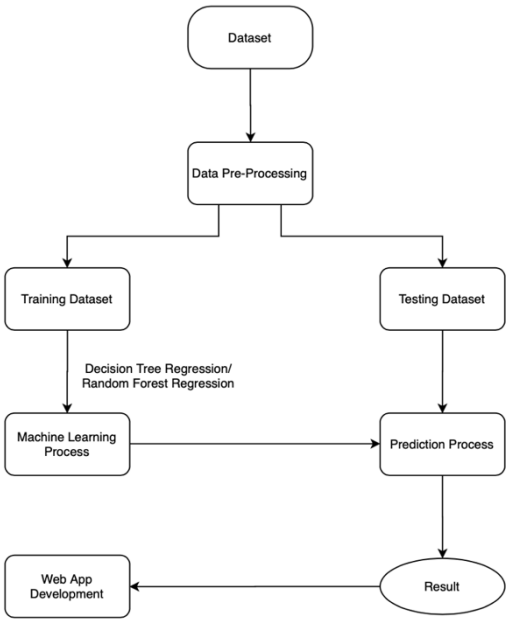
Major cause of generation of pollutants : vehicles and automobiles
Energy efficient commodities must be utilized wherever possible as they tend to save energy along with having low negative effects **[4]**.

5) Air Quality Prediction: Big Data and Machine Learning Approaches

The traditional approaches for air quality prediction use mathematical and statistical techniques.
This paper reports recent literature study, reviews and compares current research work on air quality evaluation based on big data analytics, machine learning models and techniques.
The major objective is to provide a snapshot of the vast research work and useful review on the current state-of-the-art on applicable big data approaches and machine learning techniques for air quality evaluation and predication **[5]**.

Chapter - 3: Architecture

1. Dataset
2. Data Pre-Processing
3. Model Training
4. Result
5. Web Application



In the starting of the project we research about our topic and according to relevance we find dataset from Kaggle and understand different methods to understand air quality. Then we start data pre-processing to increase efficiency of data. Then we divide data into two parts, one for training the dataset and another for testing the model. We train 4 ML models in total but we find relevance for two ML models for our project : random forest regression and decision tree regression. We check it again and again and once again do literature review. With the final outcome we made a web app to represent data analysis and prediction of new data properly using flask.

Chapter - 4: Implementation Details

1. Dataset:

The data is a combined (across the years and states) and largely clean version of the Historical Daily Ambient Air Quality Data released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP) year 1990-2015.

Data Description (features)

- stn_ code (station code)
- sampling_date (date of sample collection)
- state (Indian State)
- location (location of sample collection)
- agency
- type (type of area)
- so2 (sulphur dioxide concentration µg/m3)
- no2 (nitrogen dioxide concentration µg/m3)
- rspm (respirable suspended particulate matter concentration µg/m3)
- spm (suspended particulate matter µg/m3)
- location_monitoring_station
- pm2_5 (particulate matter 2.5 µg/m3)
- date

In the dataset there are 435742 rows and 13 columns.

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	NaN	1990-02-01
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	1990-02-01
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	1990-02-01
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	1990-03-01
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	1990-03-01

Figure 1 Dataset Preview

2. Data Pre-Processing

- There are 435742 rows and 13 columns in the dataset.
- Firstly, we will remove unwanted columns like stn_code, agency, sampling_date, location_monitoring_station, location, type, date, pm2_5.
- Then, we find that there are many NANs(Not a Number) in the dataset. Number of NANs value in the dataset spm: 237387(54.48%), rspm: 40222(9.23%), so2: 34646(7.95%), no2: 16233(3.72%).

	Null_Values	Percent
spm	237387	54.478797
rspm	40222	9.230692
so2	34646	7.951035
no2	16233	3.725370
type	5393	1.237659

Figure 2 Number of Null Values in Dataset

- There are many null values in the dataset.
- So, we will take the mean of all factors like spm, rspm, so2, no2 which are grouped by states. Then put the mean value of the factor in that state in place of null values in that particular state.

state	spm	rspm	so2	no2
Andhra Pradesh	200.260378	78.182824	7.284845	21.704451
Arunachal Pradesh	NaN	76.629213	3.179104	5.469697
Assam	153.355386	93.724912	6.723263	14.793691
Bihar	276.917416	123.705176	19.381476	36.575525
Chandigarh	206.056150	96.587079	2.676986	18.619404

Figure 3 Finding mean based on States

- By doing so we have eliminated most of the null values. Remaining number of NANs value in the dataset spm: 4071, rspm: 3, so2: 3, no2: 3.
- For removing the remaining null values, we will remove the tuples which contain null values.
- Then, we find individual particles' index of each factor according to the formula
- Add this individual particles' index in the dataset.
- Calculate AQI which is the maximum of these individual particles' index and add to the dataset.
- Now, the data is processed and ready for use.

	state	location	type	so2	no2	rspm	spm
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	78.182824	200.260378
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	78.182824	200.260378
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	78.182824	200.260378
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	78.182824	200.260378
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	78.182824	200.260378

Figure 4 Processed Dataset

3. Algorithm used to Trained Model

1. Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

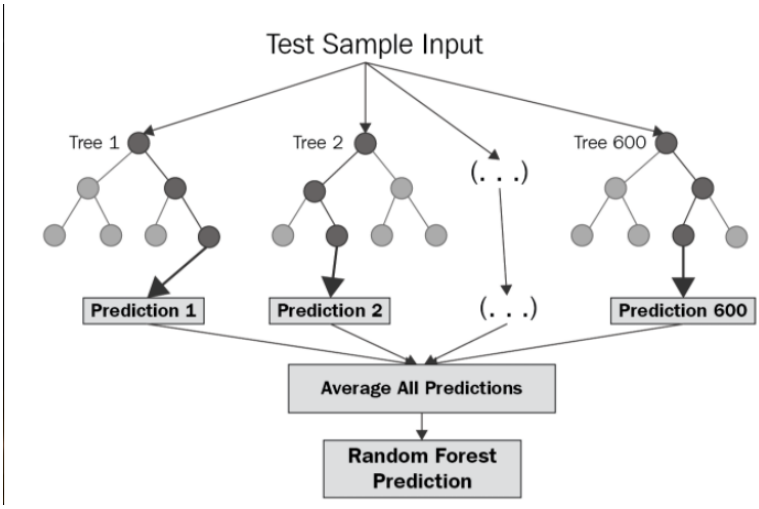


Figure 5 Random Forest Regression

The diagram above shows the structure of a Random Forest. You can notice that the trees run in parallel with no interaction amongst them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. To get a better understanding of the Random Forest algorithm, let's walk through the steps:

- Pick at random *k* data points from the training set.
- Build a decision tree associated with these *k* data points.
- Choose the number *N* of trees you want to build and repeat steps 1 and 2.
- For a new data point, make each one of your *N*-tree trees predict the value of *y* for the data point in question and assign the new data point to the average across all of the predicted *y* values.

2. Decision Tree Regression

A Decision Tree is a predictive model that uses a set of binary rules in order to calculate the dependent variable. Each tree consists of branches, nodes, and leaves. Let's familiarize ourselves with some terminology before moving forward:

- The root node represents the entire population and is divided into two or more homogeneous sets.
- A decision node is when a sub-node splits into further sub-nodes.
- A leaf is when a node does not split. *These are also referred to as "Terminal Nodes".*

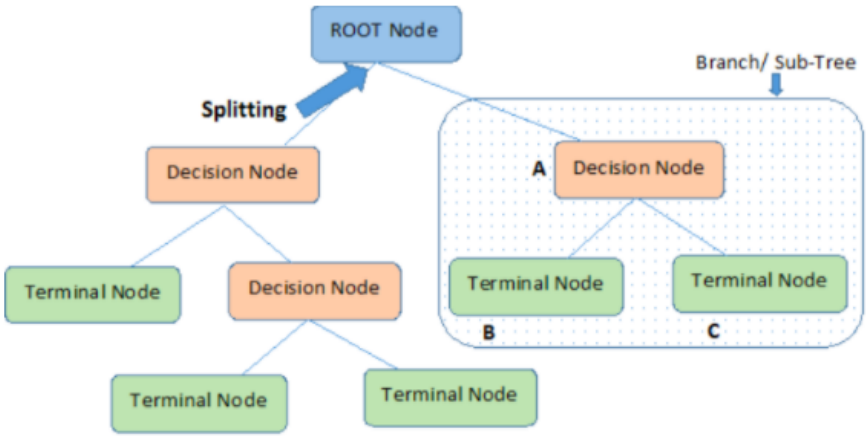


Figure 6 Decision Tree Regression

4. Data Visualization

For Data Visualization, we have use Tableau Application.

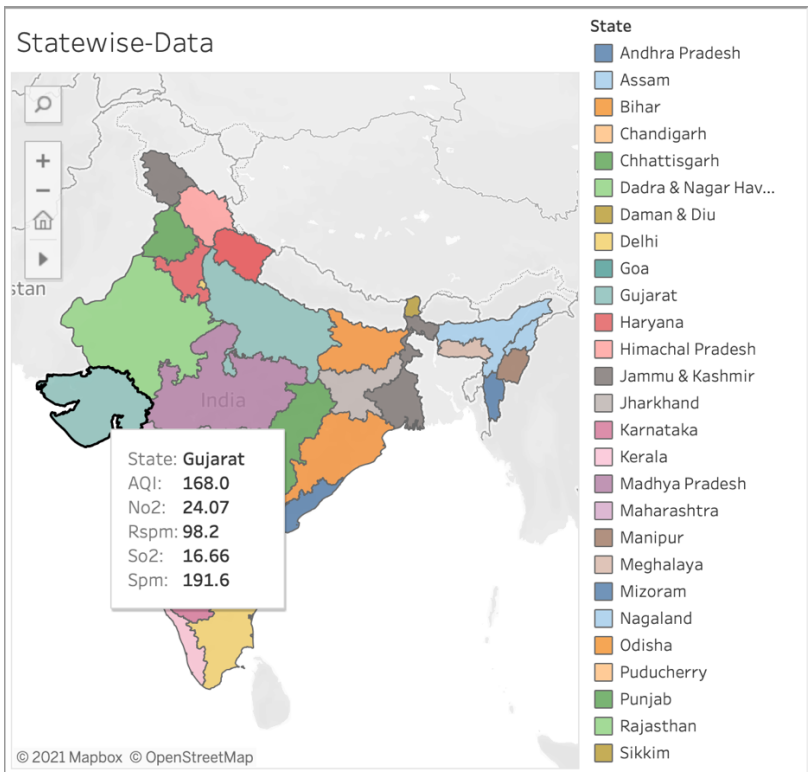


Figure 7 State-wise Data Visualization

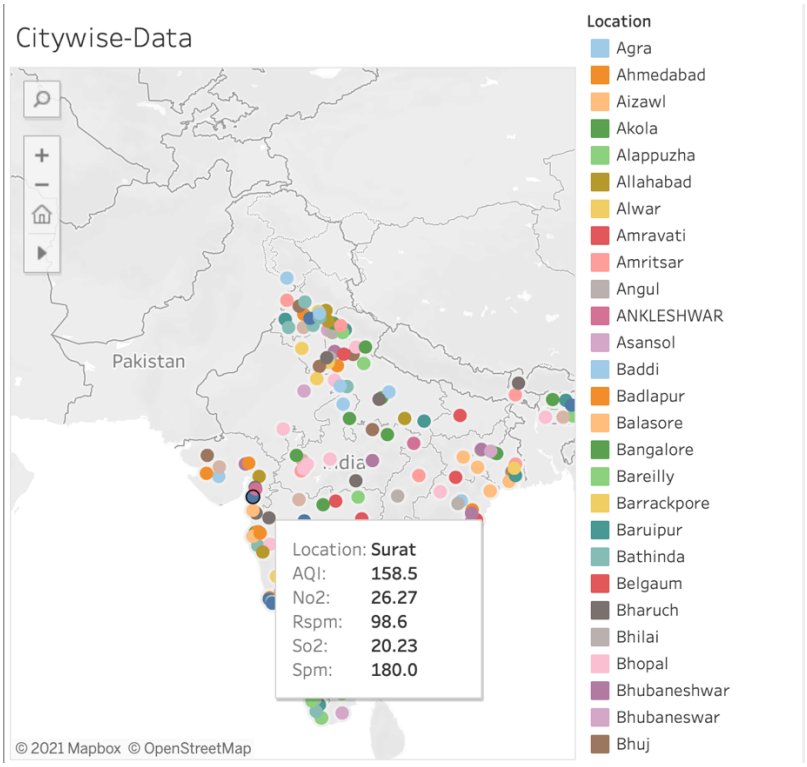


Figure 8 City-wise Data Visualization

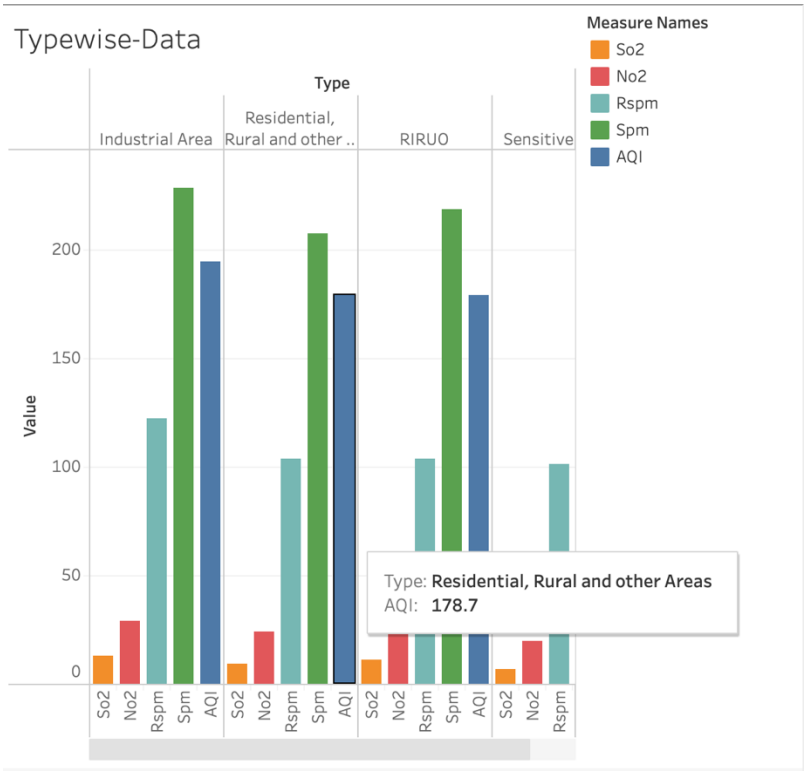


Figure 9 Type-wise Data Visualization

5. Creating Web-Application

For creating Web-App we use the Flask framework. Flask is a web framework, it's a Python module that lets you develop web applications easily. It has a small and easy-to-extend core: it's a microframework that doesn't include an ORM (Object Relational Manager) or such features. It does have many cool features like URL routing, template engine. It is a WSGI web app framework.

It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Applications that use the Flask framework include Pinterest and LinkedIn.

Features

- Development server and debugger
- Integrated support for unit testing
- RESTful request dispatching
- Uses Jinja templating
- Support for secure cookies (client side sessions)
- 100% WSGI 1.0 compliant
- Unicode-based
- Extensive documentation
- Google App Engine compatibility
- Extensions available to enhance features desired

Key difference between flask and Django

- Flask provides support for API while Django doesn't have any support for API.
- Flask does not support dynamic HTML pages and Django offers dynamic HTML pages.
- Flask is a Python web framework built for rapid development whereas Django is built for easy and simple projects.
- Flask offers a diversified working style while Django offers a Monolithic working style.
- The URL dispatcher of the Flask web framework is a RESTful request on the other hand, the URL dispatcher of Django framework is based on controller-regex.
- Flask is a WSGI framework while Django is a Full Stack Web Framework.

For the above reasons, we have used flask framework to create a Web-App of our model.

Snapshots of the Web-App are as follows:

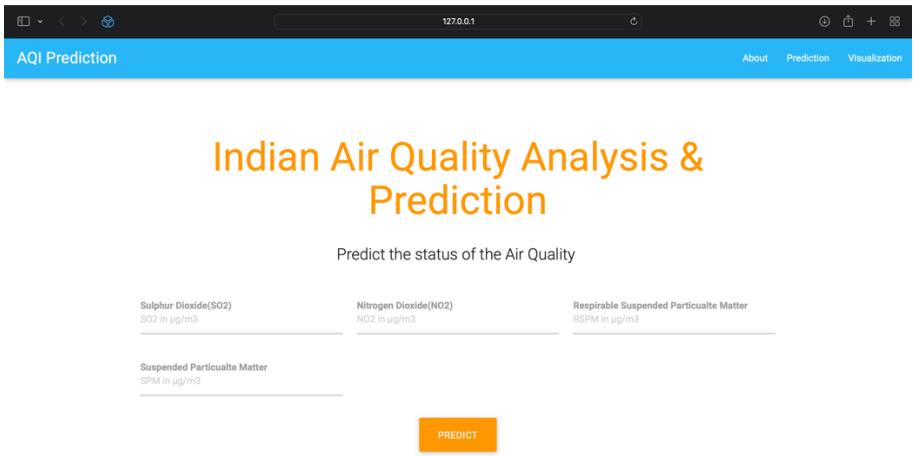


Figure 10 Home page of Web-App

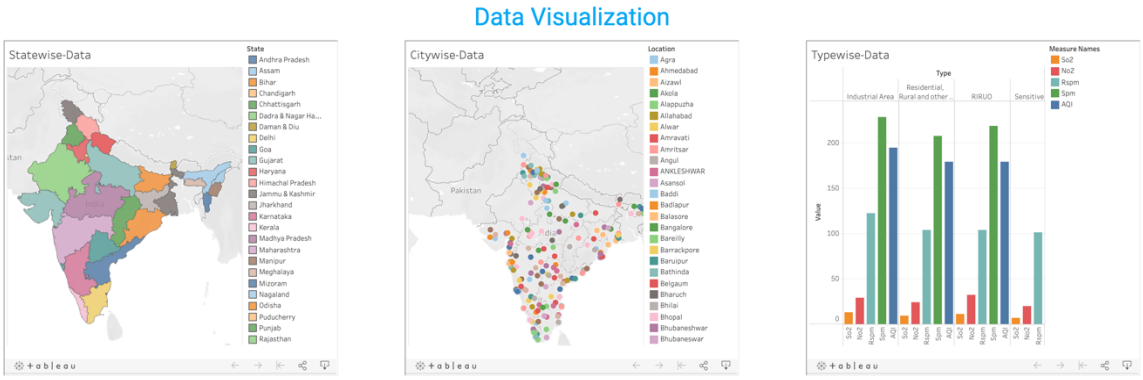


Figure 11 Data Visualization part of Web-App

Chapter - 5: Results and Comparison with Existing Work

In this section, we will discuss the performance of our trained models. We will compare the models on different metrics like Linear Regression, Logistic Regression, Random forest Regression, and Decision tree Regression.

Algorithm	Training Time(s)	Accuracy	Error(RMSE)
Linear Regression	0.15	94.65%	20.08
Logistic Regression	2.73	74.02%	54.27
Random Forest Regression	59.09	99.983%	1.14
Decision Tree Regression	0.84	99.978%	1.31

As shown in the above table, the Accuracy of Random forest Regression and Decision Tree Regression is highest but, as the Training time of Random Forest Regression is far more than Decision Tree Regression. We choose Decision Tree Regression as the preferred algorithm to train the model.

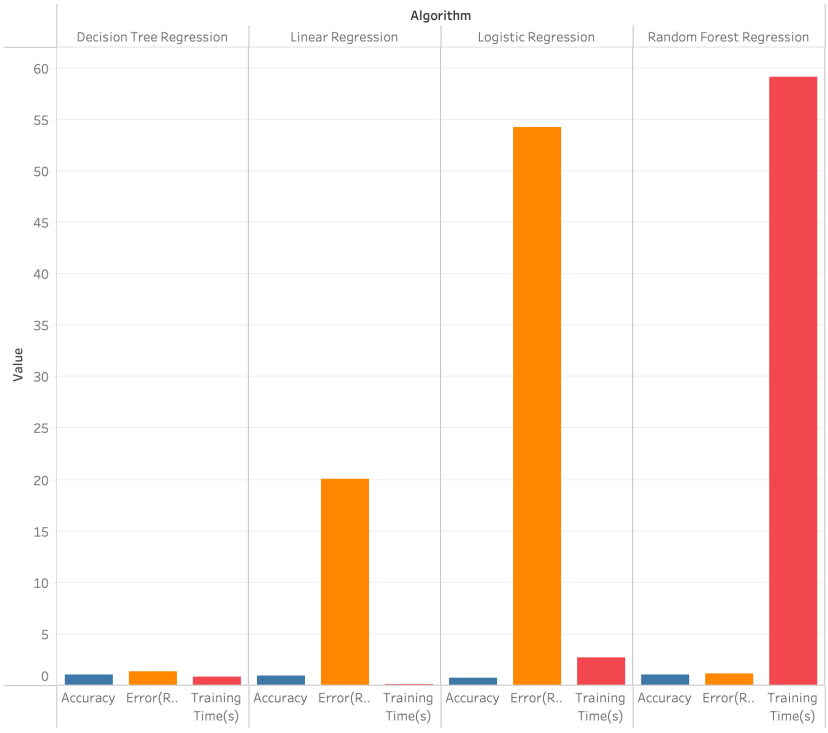


Figure 12 Comparison of various Algorithm

Chapter – 6: Summary and Future Directions

Summary

All in all entire project is about see some of the factors which create degradation of air quality and make things more visible. With this type of data analysis and prediction mechanism we also need to understand the practical reasons of pollution of air in certain part of country like open waste burning, heavy industry wastage, inefficient burning of fuel many more. For the basic awareness, government should put law for describing of pollution content on the product itself. students should run awareness campaign for saving the environment. Those kind of practical implementation and proper study based on data on real time bases. Right now we are borrowing our material wants from future generation and gives them polluted environmental conditions.

Future Directions

- Putting new updated data and studies related to air quality.
- Ground level policy implementers became more aware about how air quality depends and its dependents.
- We can also provide an e-commerce platform for protection from low air quality tools.
- NGOs, SHGs have a better chance to work with government agencies and raise funds according to data prediction.
- Industries like chemical, pharmaceutic, textile, chip etc have better idea about air quality and their impact on their industries.
- Current features can be replace by real life features like Number of Vehicles or Industries etc.

References:

1. M. Soundari, J. Jeslin and Akshaya, “INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING.” (2019).

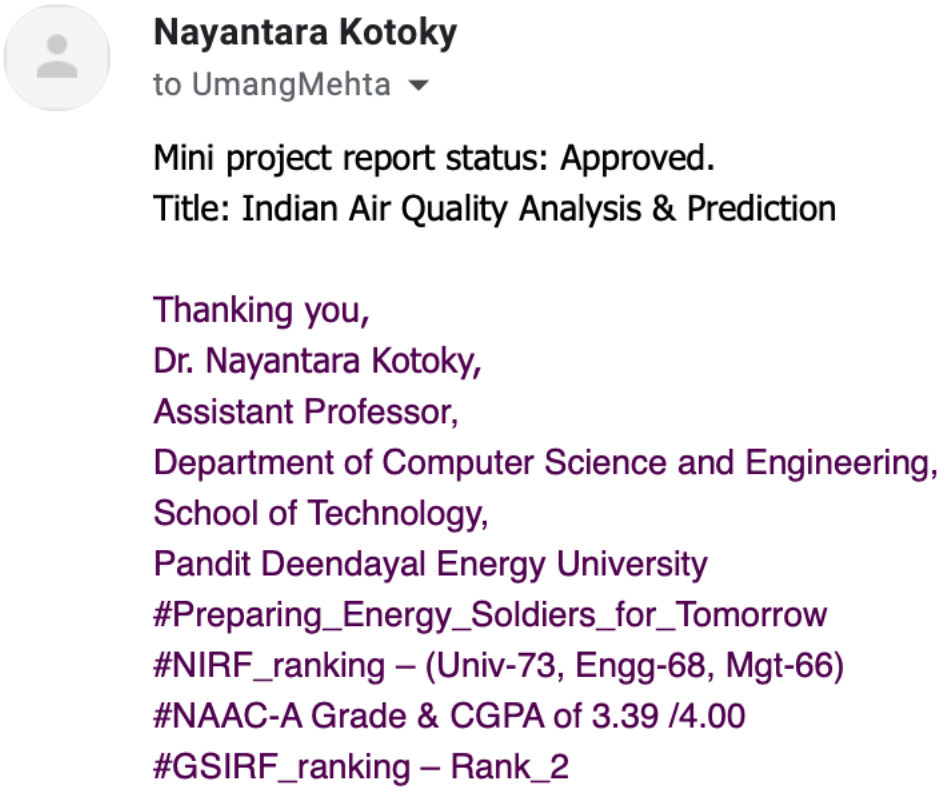
2. T. Madan, S. Sagar and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 140-145.

3. Li, Sheng-Tun & Shue, Li-Yen. (2004). Data mining to aid policy making in air pollution management. Expert Systems with Applications. 27. 331-340.

4. Nasir, Humaib & Goyal, Kirti & Prabhakar, Dolonchapa. (2016). Review of Air Quality Monitoring: Case Study of India. Indian Journal of Science and Technology.

5. Kaur, Gaganjot & Gao, Jerry & Chiao, Sen & Lu, Shengqiang & Xie, Gang. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. International Journal of Environmental Science and Development.

Mentor's Approval:



Nayantara Kotoky
to UmangMehta

Mini project report status: Approved.
Title: Indian Air Quality Analysis & Prediction

Thanking you,
Dr. Nayantara Kotoky,
Assistant Professor,
Department of Computer Science and Engineering,
School of Technology,
Pandit Deendayal Energy University
#Preparing_Energy_Soldiers_for_Tomorrow
#NIRF_ranking – (Univ-73, Engg-68, Mgt-66)
#NAAC-A Grade & CGPA of 3.39 /4.00
#GSIRF_ranking – Rank_2