A PROJECT REPORT

on

"AIR QUALITY INDEX ANALYSIS"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE & ENGINEERING

BY

ROHIT KUMAR	21052984
SHUBHAM	2105751
INDRANIL BHATTACHARJEE	21051983
ABHINEET YADAV	21051260

UNDER THE GUIDANCE OF
Ms. Priyanka Roy

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

NOVEMBER 2024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, Odisha,751024

CERTIFICATE

This is to certify that the project entitled

“AIR QUALITY INDEX ANALYSIS“

submitted by

ROHIT KUMAR	21052984
SHUBHAM	2105751
INDRANIL BHATTACHARJEE	21051983
ABHINEET YADAV	21051260

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be University, Bhubaneswar. This work is done during year 2024-2025, under our guidance.

Date: 14 / 11 / 2024

(Ms.Priyanka Roy)
Project Guide

Acknowledgment

We are profoundly grateful to **Ms. Priyanka Roy** for her expert guidance and continuous encouragement throughout to see that this project has reached its target since its commencement to its completion. ....

ROHIT KUMAR

SHUBHAM

INDRANIL BHATTACHARJEE

ABHINEET YADAV

ABSTRACT

This project aims to analyze air quality by calculating the Air Quality Index (AQI) for different regions over a specified period. The AQI is a measure used to assess air pollution levels, considering key pollutants such as PM2.5, PM10, carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). The primary goal of this project is to evaluate how pollution levels fluctuate and identify potential sources of air quality deterioration. Data was collected from publicly available air quality monitoring stations and processed to generate the AQI using the formula prescribed by the Central Pollution Control Board (CPCB).

The analysis revealed significant variations in AQI, with certain urban areas exhibiting consistently poor air quality due to high concentrations of particulate matter and industrial emissions. This report discusses the trends observed, the health implications of varying AQI levels, and possible measures to mitigate air pollution. The findings underscore the urgent need for improved air quality management and the adoption of sustainable practices to protect public health and the environment.

Contents

Chapter 1 Introduction

Air pollution is a critical issue that affects millions of people worldwide, and it is particularly severe in India, where the rapid pace of industrialization, urbanization, and population growth has led to a dramatic decline in air quality. India is home to 13 of the 20 most polluted cities in the world, with harmful air pollutants consistently exceeding safe levels set by the World Health Organization (WHO). The consequences of this widespread pollution are dire, with air pollution ranking as one of the top causes of premature deaths in the country, surpassing traditional health risks such as high blood pressure and smoking. According to recent reports, air pollution contributes to the premature death of over 2 million people in India each year.

The main sources of air pollution in India are different. In urban areas, vehicular emission and industrial processes constitute the largest concentration of pollutants in the air. Burning fossil fuels, such as coal and oil, mainly through power plants and factories, releases gases of sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) into the atmosphere.

It aims to provide insights into the Air Quality Index (AQI) and its influencing factors, such as concentrations of sulphur dioxide (SO_2), nitrogen dioxide (NO_2), suspended particulate matter (SPM), and respirable suspended particulate matter (RSPM). By leveraging algorithms like Linear Regression and Logistic Regression for regression tasks, and Decision Tree Classifier and Random Forest Classifier for classification, the project explores both prediction and classification aspects of the AQI.

Background

Presently, India faces a significant problem in the form of alarming air pollution. Industrial emissions, vehicle exhausts, and crop residue burning at agricultural levels have brought serious health concerns and environmental hazards to one fourth of its population, counting 13 of the world's most polluted cities in the country. Even though the limiting safe limits as set by the World Health Organization (WHO) are highly exceeded for many such people, millions still breathe the polluted air. PM2.5 and PM10, sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) are other particulate matter that add very significantly to the degradation of air quality. Therefore, this has resulted in an increase of respiratory diseases and cardiovascular conditions as well as premature deaths all over the country.

Monitoring air quality and accurately predicting its variations are crucial steps toward mitigating the adverse effects of pollution. The Air Quality Index (AQI) is a standardized measure that simplifies the complexity of air pollution levels, translating it into categories such as "Good," "Moderate," and "Severe." This allows the public and authorities to better understand air pollution's immediate effects and implement timely interventions.

Objective

The primary objective of this project is to analyze India's air quality data by exploring the distribution and trends of pollutants such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), suspended particulate matter (SPM), and respirable suspended particulate matter (RSPM). The project aims to predict the Air Quality Index (AQI) using regression models and classify air quality into categories such as good, satisfactory, moderately polluted, poor, very poor, and severe through classification algorithms.

Chapter 2 Literature Review

Literature Review on Air Quality and AQI

Air quality is one of the most vital environmental factors directly linked with human health, ecosystems, and global climate. Monitoring air quality is one of the easy ways of knowing the levels of pollution and all the possible impacts it might pose. The Air Quality Index, also known as AQI, is an international standardized system that allows countries around the world to compare their air-quality information. It is an aggregate measure, in terms of concentration, of a number of common air pollutants, and its application varies by area.

Calculations to determine AQI include the concentration of pollutants such as Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and Suspended Particulate Matter (SPM) with Respirable Suspended Particulate Matter (RSPM), among others. The higher the AQI value, the worse the quality of air will be, and it is more likely that adverse health effects may occur.

2.1 Global Studies and Research on Air Quality and AQI

Several studies have examined the relationship between air quality and public health, focusing on both the health impacts of pollution and the effectiveness of AQI systems in conveying risks.

- **Health Impacts of Air Pollution**:

The World Health Organization reports that in the major risk factors for respiratory and cardiovascular diseases, high exposure to air pollution, especially particulate matter fine PM2.5, have been known to claim millions of early deaths worldwide each year. There are many studies, along with scientific evidence, that linked pollutants such as PM2.5 and Ozone to exacerbations of chronic conditions like asthma, lung cancer, and heart disease.

- **Comparison of AQI Systems**:

Studies on comparative analysis of AQI systems in different countries reveal a direct contrast between the measurements and communications in place. For instance, India's AQI system is different from that of the US regarding thresholds of pollutants, categories, and health advisories. A study in comparison between the two systems of the U.S. and India AQI found that, although both tend to deliver a clean health message, the India AQI system is specifically designed to help confront specific sources of pollution, namely crop burning, industrial emission, and vehicular exhaust.

2.2 AQI Measurement Methods

The AQI is measured using a variety of methods, including traditional ground-based monitoring, satellite-based sensors, and low-cost air quality sensors. These methods provide valuable data, but each has its advantages and challenges. Ground-based air quality monitoring stations, in short, AQMS, are an indispensable tool for tracking the trends of air pollution levels and assessing health at both urban and rural environments. Stations of such in various regions -urban centers, industrial zones, and others-rather are located around the areas to collect data on concentration pollutants in the air. The data from the different stations is then processed to compute indices such as the Air Quality Index (AQI) to establish the relationship between pollutant levels and public health outcomes.

2.3 Global Standards for AQI

Different countries have developed their own AQI systems to inform the public about air quality. These systems vary in terms of the pollutants they monitor and the concentration thresholds they use. The Air Quality Index or AQI is a scale that represents the level of pollution in the air and its health effects on an individual. This standardized scale, which has been adopted by the United States, is used to inform the public about air quality. This is a simple, color-coded scale.

2.4 Challenges in AQI Measurement and Standardization

Despite the widespread use of AQI systems, several challenges remain:

- **Accuracy of Data**: Low-cost sensors may struggle with accuracy, which can affect the reliability of AQI measurements. Calibration and sensor placement are critical issues.
- **Regional Variability**: Pollution sources vary significantly between regions. In India, for example, agricultural practices like stubble burning significantly affect air quality, contributing to high levels of RSPM during certain seasons.

Chapter 3 Methodology

3.1 Data Collection

The available public data was retrieved from the Kaggle platform, India Air Quality Data. The data set has measurements of air quality done at different stations in India for multiple years. The data involves elaborate records of the main air pollutants: nitrogen dioxide (NO_2), sulfur dioxide (SO_2), respirable suspended particulate matter (RSPM), and suspended particulate matter (SPM), which are significant indicators of risks concerning air pollution and public health.

3.2 Parameters Considered

Several key air quality parameters were evaluated in determining the state of air pollution. The parameters represent critical indicators of environmental and health impacts of air quality, which help identify the nature of pollutants that most affect public health and the ecosystem. Below is a detailed explanation of each of the parameters included in the dataset:

3.2.1 SO_2 (Sulphur Dioxide)

Sulphur dioxide (SO_2) is a colourless pungent gas mainly obtained by the burning of fossil fuels containing sulphur, which may be coals, oils. Some other industrial activities, such as smelting of metals and petroleum refining, also yield this gas. It is one of the major air pollutants that has gross impacts both on human health and the environment.

3.2.2 NO_2 (Nitrogen Dioxide)

Nitrogen dioxide (NO_2) is a reddish-brown gas that forms primarily from the combustion of fossil fuels, especially in vehicles, power plants, and industrial facilities. It is a key component of nitrogen oxides (NO_x), which play a significant role in air pollution.

3.2.3 RSPM (Respirable Suspended Particulate Matter)

RSPM or PM2.5 stands for respirable suspended particulate matter; the particle matter measures 2.5 micrometers or smaller, which are very tiny and can easily be inhaled deep into the lungs. RSPM causes serious damage to the respiratory system because it may penetrate deep into the lungs and come into direct contact with the blood.

3.2.4 SPM (Suspended Particulate Matter)

Suspended Particulate Matter, for the most part, refers to the mixture of liquid droplets and solid particles suspended in the air. This category includes particles ranging from those larger than PM2.5 but smaller than 10 micrometers. SPM is used mostly as a general term referring to particulate matter in the air and not necessarily by size.

3.3 Tools & Techniques

3.3.1 Python Libraries and Tools Used

- **Pandas**

Pandas, although a powerful library used in Python for the manipulation and analysis of data, is still a highly important requirement for loading, cleaning, and most importantly, exploring your dataset. This was used in this project to: Load Data Import the dataset (CSV, Excel, etc.) into DataFrame.

- **NumPy**

NumPy is used for numerical operations, especially when working with large datasets and performing mathematical computations. It helps in handling arrays, matrices, and performing basic statistical analysis.

- **Matplotlib and Seaborn**

The two libraries are crucial for visualization purposes, where one can examine the relations among the variables as well as the distribution of data.
Matplotlib - It is useful for basic plots such as histograms, line plots, and scatter plots.
Seaborn - More advanced for heatmaps, pair plots, and box plots.

- **Scikit-learn**

Scikit-learn is the Python library of choice for machine learning, offering a natural toolbox that readily provides simple and efficient tools for data mining and analysis, especially when related to supervised learning algorithms.

3.3.2 Techniques for Data Analysis

- **Exploratory Data Analysis (EDA)**

EDA is the first step in the analysis process, the detailed understanding of the structure and nature of the dataset. Such elements include:Descriptive Statistics: Summarizing key statistics, mean, median, variance, etc. for each feature.
Correlation analysis: That correlation differentiates various types of pollutants and AQI with each other as well as among different types of pollutants. It is represented in the form of color heatmap that helps spot areas with strong relationship.

- **Feature Engineering**

Feature engineering is the process by which raw data is converted into a set of features which can be more easily used for ML. In this project:Classification of AQI: AQI was categorized in different levels (such as good, satisfactory, moderately polluted, poor, very poor, etc.) based upon their numerical value.

Missing Values Handling: missing values in pollutant variables were either imputed as in replacing the missing values with mean or median, respectively, or entire rows were deleted if the data exceeded the threshold.

3.3.3 Supervised Learning Algorithms & Model Evaluation

- **Linear Regression:**

Linear regression is undoubtedly one of the most powerful algorithms developed within statistical and machine learning, applied to modeling the relationship between a dependent variable also called target or outcome, and one or more independent variables also called predictors or features.

Why is Linear Regression Used?

Linear regression is primarily used for the following purposes:

The main application of linear regression is predicting or forecasting a continuous target variable based on one or more independent variables. Example: predict a house price using size, number of rooms, and location; predict stock price based on past performance.

Advantages of Linear Regression:

Linear regression is actually quite conceptually easy and straightforward. The mathematical formulation is not difficult to grasp, based on the hypothesis that there exists a linear relationship between the dependent and independent variables. It is computationally efficient, requiring minimal resources to train, even with large datasets. This makes it an attractive option when dealing with time-sensitive applications or limited computational power.

Disadvantages of Linear Regression:

Possibly the biggest limitation of linear regression is the assumption that the relationship between the variables under study is linear. Many real datasets contain relationships among variables that are inherently non-linear, and it is all but natural that employing linear regression in such cases may lead to an unsatisfying poor performance of the model.

Linear Regression Model 1:

Training Features: SOi, NOi, RSPMi, SPMi (calculated pollution indices)

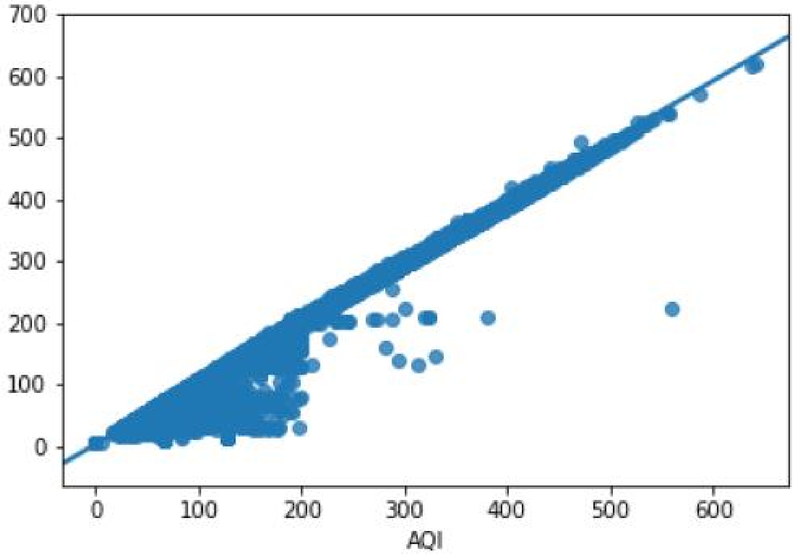
Target Feature: Calculated Air Quality Index, AQI

The data was then split as training and testing data, in which 80% of the total dataset was utilized as training data and 20% of the total dataset was utilized for testing data.

Result:

Intercept	4.0843	
Coefficients	***SOi***	0.0293
	NOi	0.0668
	RSPMi	0.0502
	SPMi	0.9549
Accuracy Score	0.9779 or 97.79 %	
R^2	0.98	
Mean Squared Error	12.89	

Regression Plot of actual values vs predicted values

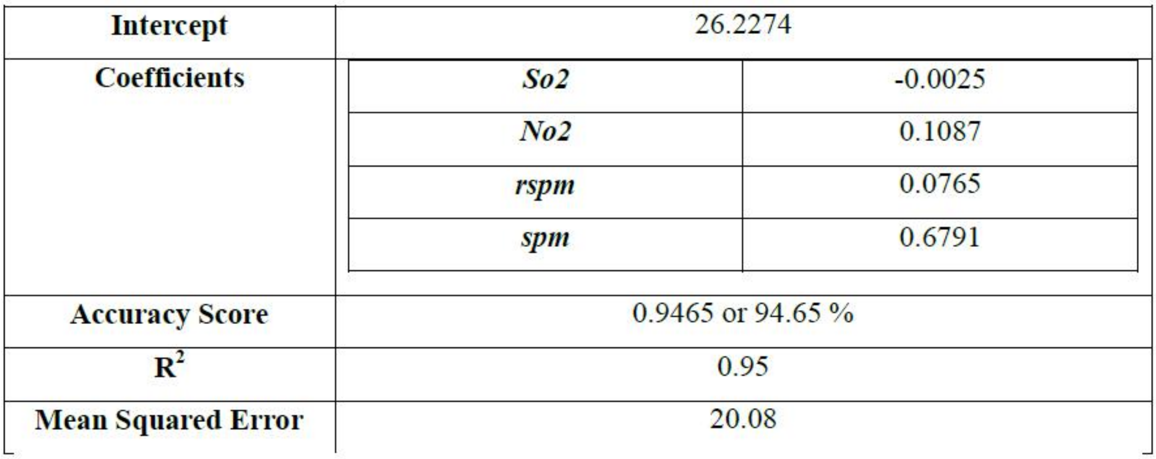


We can see that the model has a high accuracy and is a good model for AQI prediction given the value of individual Pollution indices.

Linear Regression Model 2:

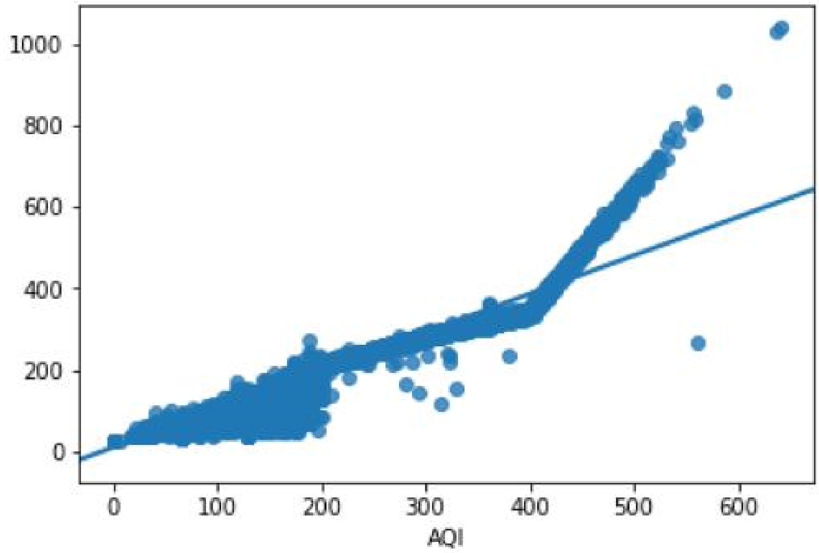
Training Features: SO_2, NO_2, RSPM, SPM (given pollutant concentration)
Target Feature: AQI (calculated Air Quality Index)
The data was split into training and testing data, where training data used 80% of the dataset and testing data used 20% of the total dataset.

Results :



Intercept	26.2274	
Coefficients	***So2***	-0.0025
	No2	0.1087
	rspm	0.0765
	spm	0.6791
Accuracy Score	0.9465 or 94.65 %	
R^2	0.95	
Mean Squared Error	20.08	

Regression Plot of actual values vs predicted values



The high accuracy value of the model shows that it is a good model for predicting AQI, given the value of individual Pollution indices; however, the Linear Regression Model 1 is superior to the second model.

➢ Logistic Regression:

Logistic regression is an analysis technique, which uses a dataset that contains one or more independent variables that might predict an outcome. The outcome is measured with a dichotomous variable; that is, there are only two possible outcomes.

Why Logistic Regression Is Used?

Logistic regression is primarily used in classification problems to classify an observation into some class based on input features. This method is essentially used when the target variable is categorized.

Advantages of Logistic Regression

Logistic regression is pretty straightforward. It's lightly computationally resource-hungry and can be quickly run even on the size of a moderately-sized dataset. That's why it's the most widely used algorithm for any binary classification task. Models are easy to interpret. The logistic regression model coefficients represent a change in the log of odds of the target variable for a one-unit change in the predictor variable.

Disadvantages of Logistic Regression

It assumes that the log-odds of the target variable and the independent variables have a linear relationship. In case the relationship is non-linear and logistic regression, without some form of transformations or a lot of feature engineering, provides not such a precise model. Like other regression techniques, logistic regression is sensitive to outliers. If there are extreme values in the dataset, it can disproportionately influence the model's estimates.

Logistic Regression Model 1:
Training Features: SO_2, NO_2, RSPM, SPM - computed pollution indices
Target Feature: AQI_Range (estimated Air Quality Index Range)
The data was divided into train and test, where training data used 80% of the total dataset and testing data used 20% of the total dataset.
Accuracy Score 77.75%
So, Logistic Regression Model 1 is not a very good classification model for this case, as the accuracy score is so low.

Logistic Regression Model 2:
Training Features: SO_2, NO_2, RSPM, SPM (given pollutant concentration)
Feature Target : AQI_Range: Calculated Air Quality Index Range.

The dataset was split into training data and testing data, with the training data using 80% of the dataset and the testing data using 20% of the dataset in total.

Accuracy Score 74.02%

So, Logistic Regression Model 2 is not a very good classification model for this case, as the accuracy score is so low.

- **Random Forest Classifier Model:**

Random forests, or sometimes random decision forests, are an ensemble learning method for classification, regression, and many other tasks that work by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

Why is Random Forest Used?

This algorithm is used for both classification and regression problems. In fact, random forest is one of the strongest and most flexible machine learning algorithms - widely used to address real-world applications in many kinds of areas, including finance, healthcare, marketing, and more.

Advantages of Random Forest Classifier

Random Forest is known for high accuracy and good generalization. It avoids the overfitting problem to some extent that one single decision tree suffers from, and this ensemble approach, using many trees, makes it a very reliable model for classification and regression tasks.

Disadvantages of Random Forest Classifier

The major weakness is that Random Forest is not interpretable compared to a simpler model such as logistic regression or a single decision tree. Although you can get the feature importance and see the decision trees, it's hard to understand the whole process of decision-making in the entire forest. This may be a huge problem in applications where model transparency is a requirement, like in healthcare or finance.

Training Features: SO_2, NO_2, RSPM, SPM (concentration of the pollutant given)

Target Feature: AQI_Range (Computed Air Quality Index Range)

The given data was divided into training and testing data. In our case, 80% of the dataset was used for training data and 20% of the total dataset was used for testing data.

Accuracy Score 99.67%

From the results above, we can see that this Random Forest Classifier model is a very good classification model in this case because of its higher accuracy score.

➢ Decision Tree Classifier Model:

One of the most simple and wide usage classifications techniques is Decision Tree Classifier. It introduces a relatively straightforward concept to solve the problem of classification. Decision Tree Classifier presents a series of astutely designed questions pertinent to the attributes of the test record. It asks the next question when it receives an answer every time until it comes to a conclusion regarding the class label of the record.

Why is Decision Tree Classifier Used?

Decision Trees are primarily used for classification tasks. For example, a Decision Tree might classify whether an email is spam or not spam, whether a customer will churn or not churn, or whether a patient has a certain disease or not based on the input features.

Advantages of Decision Tree Classifier

The most important benefit of Decision Trees is that they are intrinsically intuitive and graphical. One can visually see how a decision was made by tracing the path up the branches of the tree from the root up to the leaf node. This is a necessary factor when there is a need for explainability, such as in areas like healthcare and finance.

Disadvantages of Decision Tree Classifier

Decision Trees suffer from a very high probability of overfitting, especially if the tree is deep, that is, has too many levels. If no form of pruning or regularization is applied to the tree, it is likely to memorize the training data and that directly translates into poor generalization performance on unseen data.

Training Features: SO_2, NO_2, RSPM, SPM (given pollutant concentration)
Target feature AQI_Range = calculated Air Quality Index Range
It was divided into two parts: one for training data and the other for testing data, where the training data is the one that used 80% of the dataset, while the testing data used 20% of the total dataset.

Accuracy Score 99.98%

We can find that the Decision Tree Classifier Model is the best for this classification since it has a high accuracy score.

Model	**Accuracy Score(in percentage)**
Linear Regression Model 1	97.79
Linear Regression Model 2	94.65
Logistic Regression Model 1	77.75
Logistic Regression Model 2	74.02
Random Forest Classifier	99.67
Decision Tree Classifier	99.98

Chapter 4 Data Analysis

4.1 Dataset

Data version of the Historical combined, across the years and states, and largely clean Daily Ambient Air Quality Data published by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Portal has set up a data portal at its website: www.cpcb.nic.in central-pollution-board.

4.2 Data Description (features)

1. Station code-(stn_code)
2. Date of sample collection-(sampling_date)
3. Indian State-(state)
4. Location of sample collection-(location)
5. Agency
6. Type of area-(type)
7. Sulphur dioxide-(SO_2)
8. Nitrogen dioxide-(NO_2)
9. Respirable suspended particulate matter-(RSPM)
10. Suspended particulate matter-(SPM)
11. Location monitoring station
12. Particulate matter-(pm2_5)
13. Date

To effectively process raw data, the approach typically involves a series of steps aimed at transforming the unstructured or semi-structured data into a clean, structured, and analyzable format. Here's a general outline of the steps:

1. **Data Collection**: Gathered the raw data from Kaggle.
2. **Data Inspection**: Examine the raw data to understand its structure, quality, and any missing or anomalous values. This helps define the steps for cleaning and transforming the data.
3. **Data Cleaning**:
    - **Handling Missing Values**: Depending on the context, missing values can be removed, filled with averages/medians, or interpolated.
    - **Removing Duplicates**: Identifying and removing duplicate records to avoid bias.
    - **Outlier Detection and Treatment**: Outliers are identified and either removed or adjusted to avoid skewing analyses.
4. **Data Transformation**:
    - Normalisation/Scaling: As per the nature of the data, normalization, also known as rescaling the data to a 0-1 range, or standardization, rescaling the data in a way to have zero mean and unit variance could be applied.

- Encode categorical variables: Transform categorical variables into some format that can be used for computations, such as through one-hot encoding, or label encoding.
- Feature Engineering: It describes the creation of new features or transformation of existing features to make it capture better whatever information might be relevant for the analysis or model.

5. **Data Integration** (if needed): It takes data from multiple sources and forms a single dataset by matching and combining common keys or attributes.

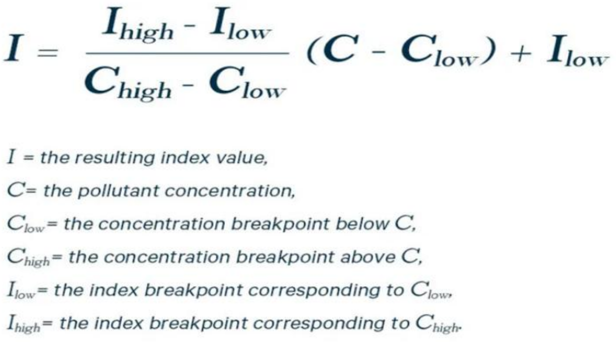
The AQI calculation involves the following steps:

1. Identify AQI Breakpoints

For each pollutant, different concentration ranges (known as breakpoints) correspond to AQI values. For example, the AQI for $PM_{2.5}$ might have a breakpoint where a concentration between 12.1 and 35.4 μg/m^3 corresponds to AQI values from 51 to 100. These breakpoints vary by pollutant.

2. Apply the AQI Formula

For a pollutant with concentration CCC, its AQI III can be calculated using linear interpolation with the following formula:



$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}} (C - C_{low}) + I_{low}$$

I = the resulting index value,
C= the pollutant concentration,
C_{low}= the concentration breakpoint below C,
C_{high}= the concentration breakpoint above C,
I_{low}= the index breakpoint corresponding to C_{low},
I_{high}= the index breakpoint corresponding to C_{high}.

3. Calculate AQI for Each Pollutant

Each pollutant's concentration is used to calculate its AQI individually. This yields a separate AQI value for each pollutant.

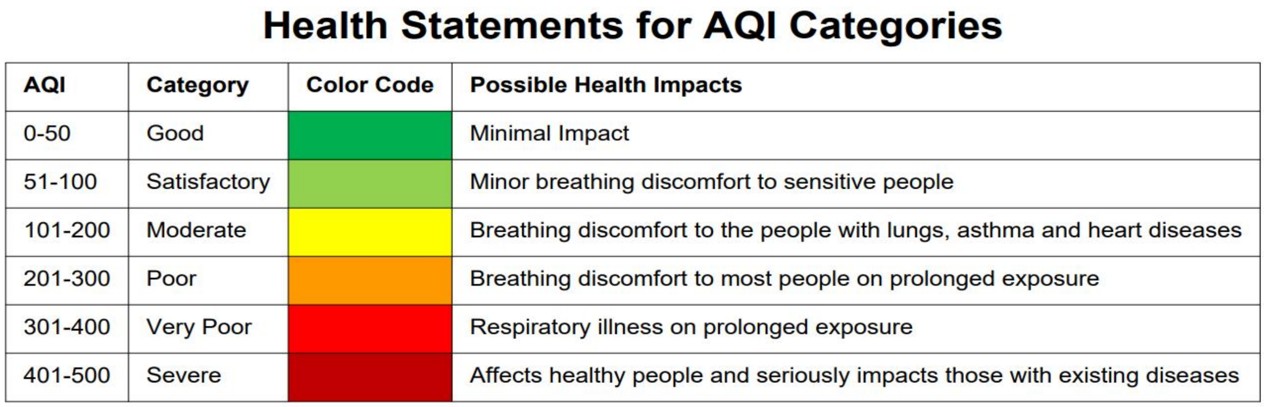
4. Determine the Overall AQI

The overall AQI for the location is the highest AQI among all the pollutants, known as the *dominant pollutant*. For example, if the AQI values are 80 for ozone, 120 for $PM_{2.5}$, and 60 for SO_2, the overall AQI is 120.

4.3 AQI Categories

AQI values are divided into categories based on health impact:

- 0–50: Good
- 51–100: Moderate
- 101–150: Unhealthy for Sensitive Groups
- 151–200: Unhealthy
- 201–300: Very Unhealthy
- 301–500: Hazardous



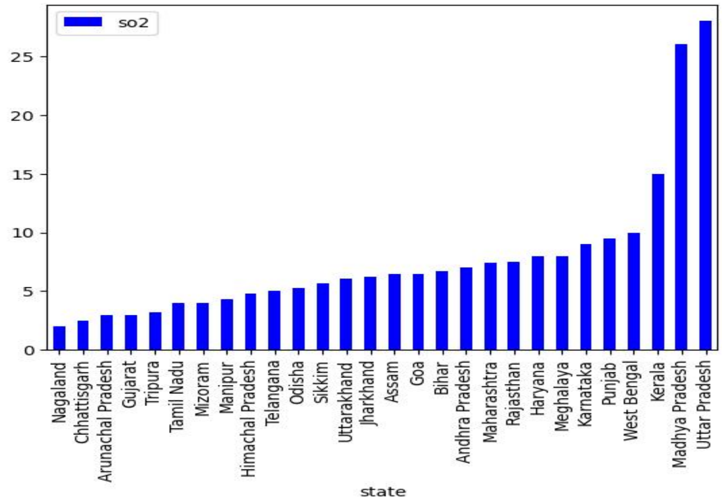
Health Statements for AQI Categories

AQI	Category	Color Code	Possible Health Impacts
0-50	Good	🟩	Minimal Impact
51-100	Satisfactory	🟩	Minor breathing discomfort to sensitive people
101-200	Moderate	🟨	Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor	🟧	Breathing discomfort to most people on prolonged exposure
301-400	Very Poor	🟥	Respiratory illness on prolonged exposure
401-500	Severe	🟥	Affects healthy people and seriously impacts those with existing diseases

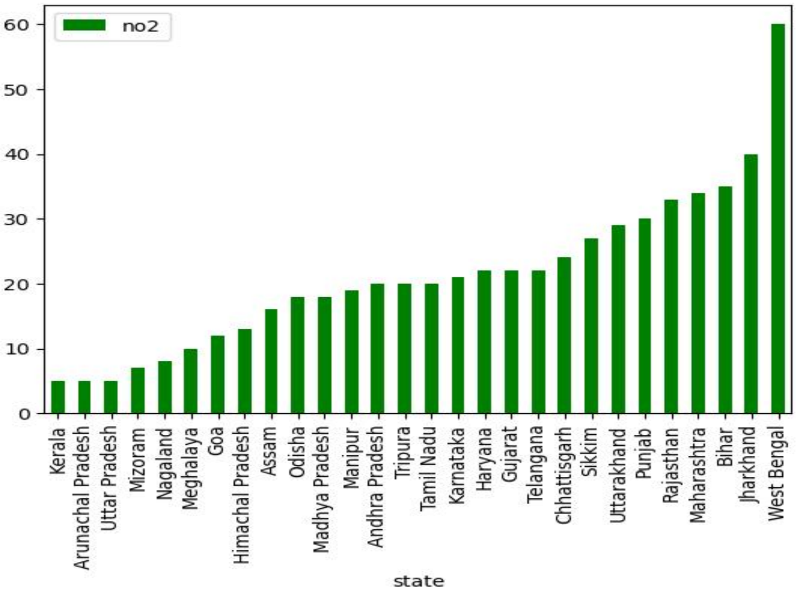
This scale makes it easy for the public to understand the health implications of local air quality.To visualize AQI levels effectively over time and by location, you can use a range of graphs, charts, and tables that highlight trends, patterns, and comparisons. Here are some visualization options that can provide clear insights into the AQI data.

4.4 Visualization of States and individual Pollutants

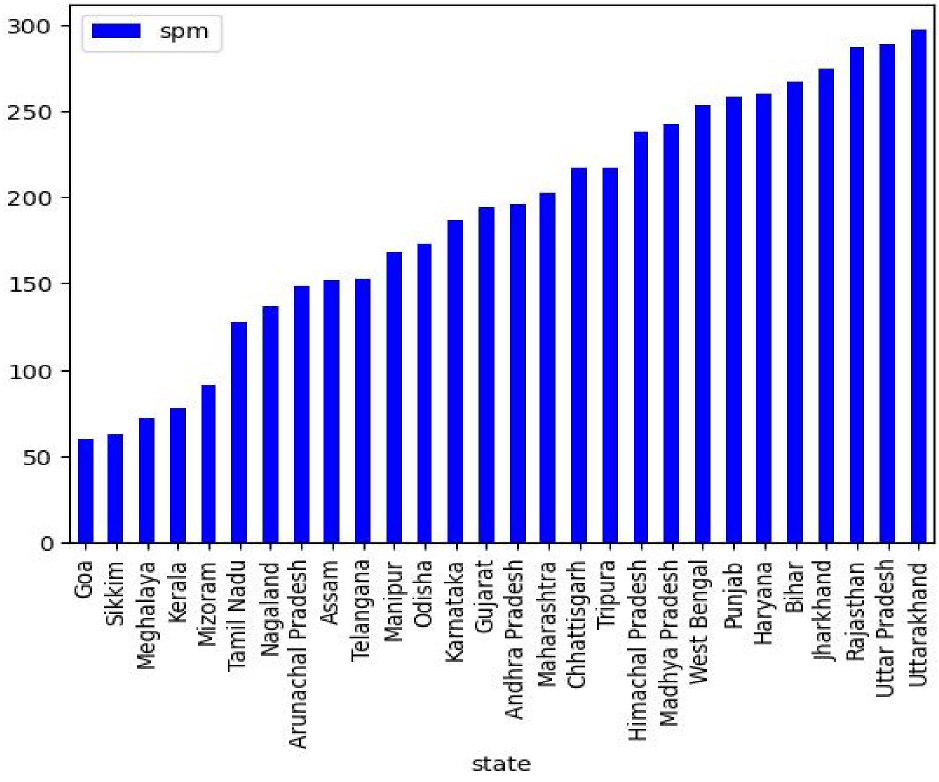
1)Sulphur Dioxide Concentration



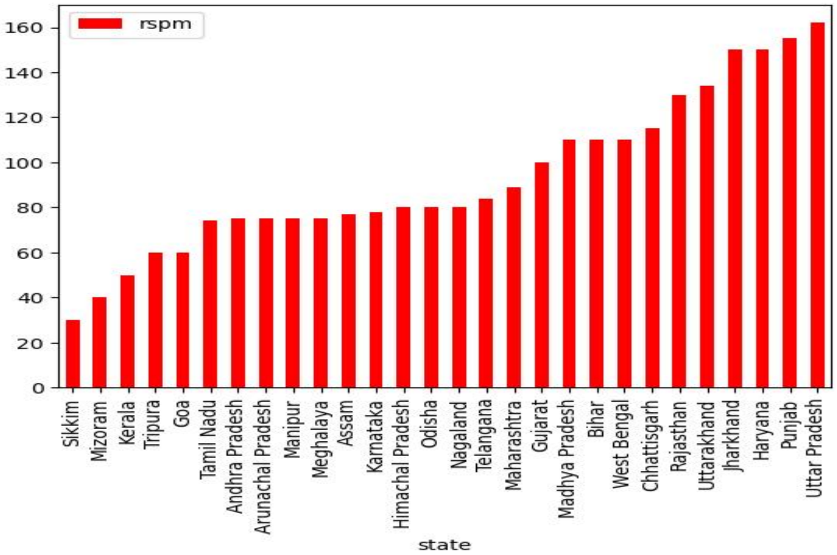
2)Nitrogen Dioxide Concentration



3)Suspended Particulate Matter Concentration

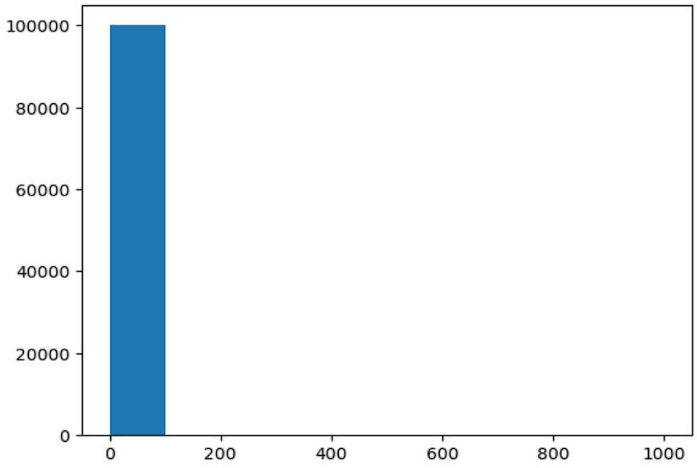


4) Respirable Suspended Particulate Matter Concentration

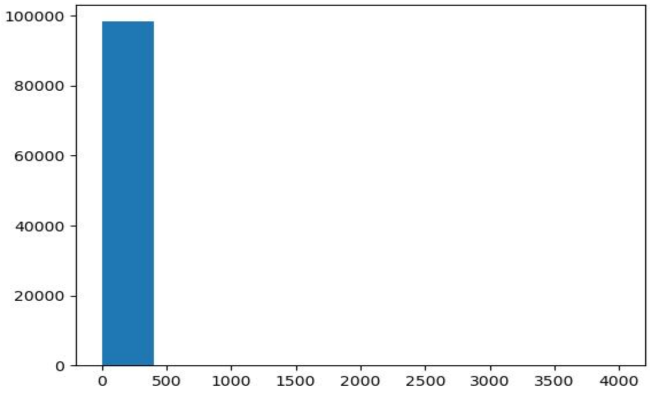


Data Distribution:

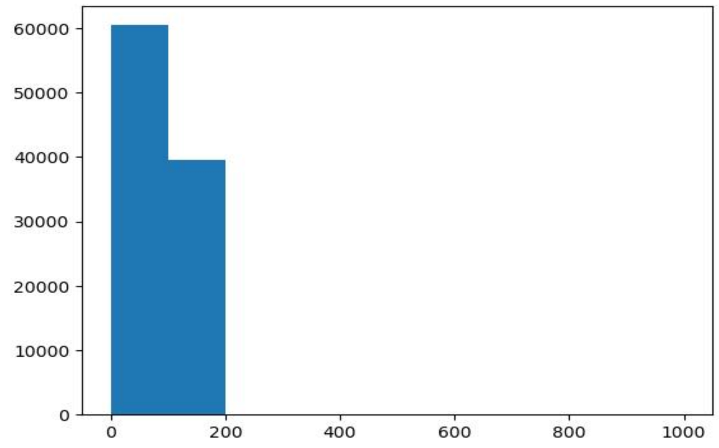
1)Sulphur Dioxide Concentration



2)Suspended Particulate Matter

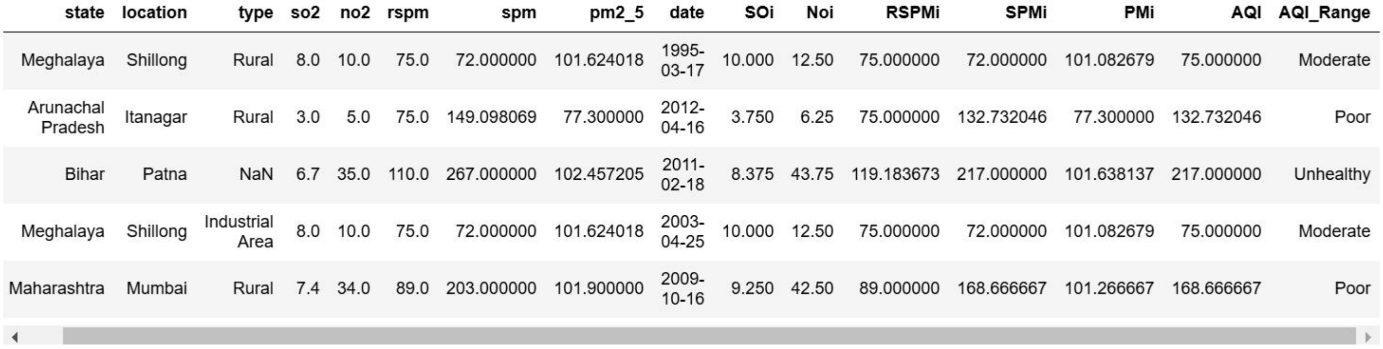


4)Respirable Suspended Particulate Matter



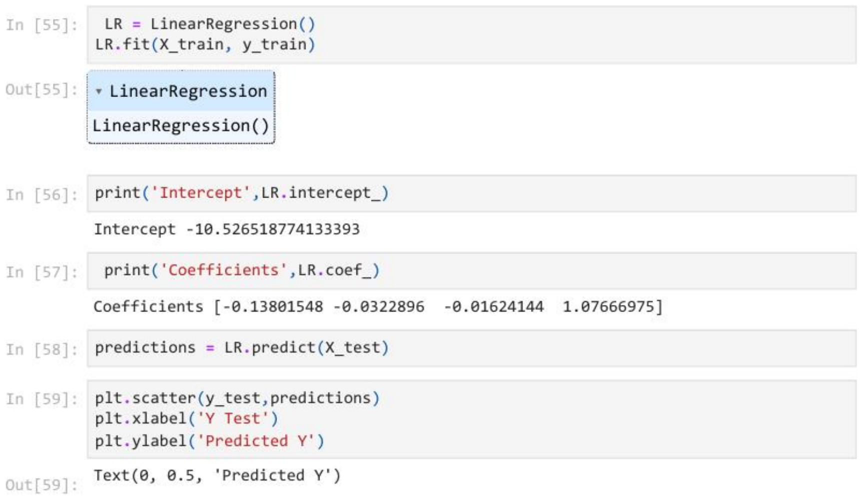
Chapter 5 Result

AQI and AQI Range of different state

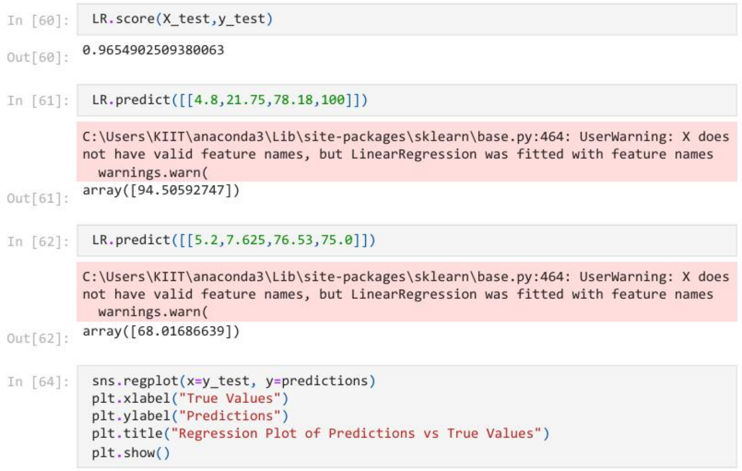


state	location	type	so2	no2	rspm	spm	pm2_5	date	SOi	Noi	RSPMi	SPMi	PMi	AQI	AQI_Range
Meghalaya	Shillong	Rural	8.0	10.0	75.0	72.000000	101.624018	1995-03-17	10.000	12.50	75.000000	72.000000	101.082679	75.000000	Moderate
Arunachal Pradesh	Itanagar	Rural	3.0	5.0	75.0	149.098069	77.300000	2012-04-16	3.750	6.25	75.000000	132.732046	77.300000	132.732046	Poor
Bihar	Patna	NaN	6.7	35.0	110.0	267.000000	102.457205	2011-02-18	8.375	43.75	119.183673	217.000000	101.638137	217.000000	Unhealthy
Meghalaya	Shillong	Industrial Area	8.0	10.0	75.0	72.000000	101.624018	2003-04-25	10.000	12.50	75.000000	72.000000	101.082679	75.000000	Moderate
Maharashtra	Mumbai	Rural	7.4	34.0	89.0	203.000000	101.900000	2009-10-16	9.250	42.50	89.000000	168.666667	101.266667	168.666667	Poor

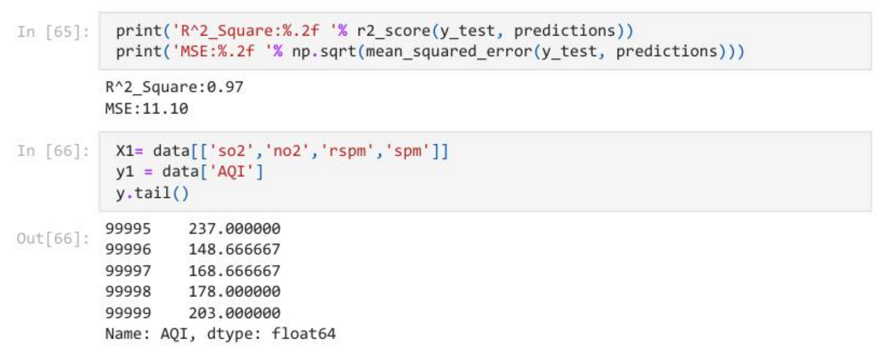
Algorithms and their Implementation



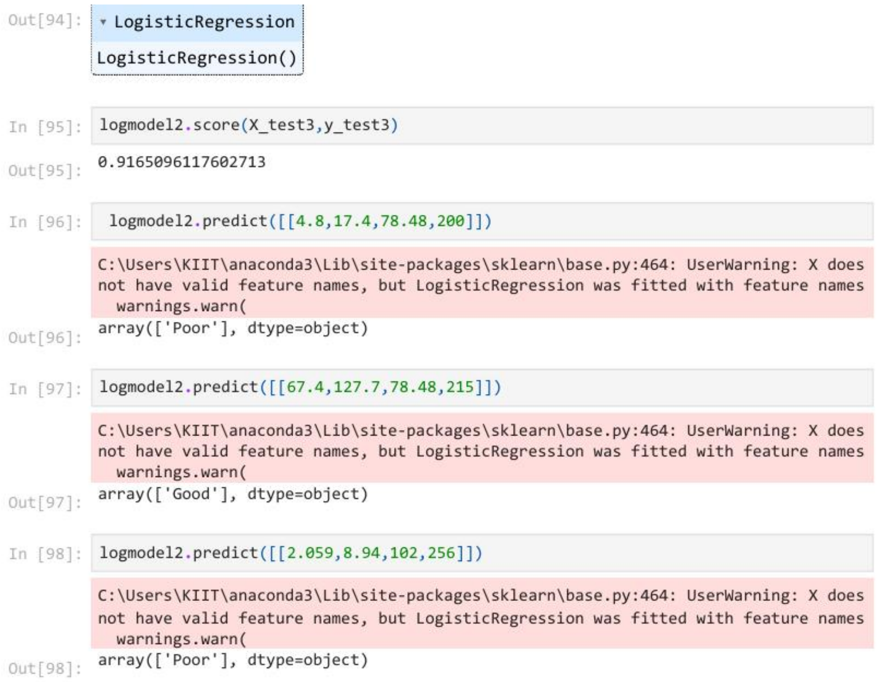
```
In [55]:  LR = LinearRegression()
          LR.fit(X_train, y_train)

Out[55]: ▾ LinearRegression
          LinearRegression()

In [56]: print('Intercept',LR.intercept_)

         Intercept -10.526518774133393

In [57]:  print('Coefficients',LR.coef_)

         Coefficients [-0.13801548 -0.0322896  -0.01624144  1.07666975]

In [58]: predictions = LR.predict(X_test)

In [59]: plt.scatter(y_test,predictions)
         plt.xlabel('Y Test')
         plt.ylabel('Predicted Y')

Out[59]: Text(0, 0.5, 'Predicted Y')
```



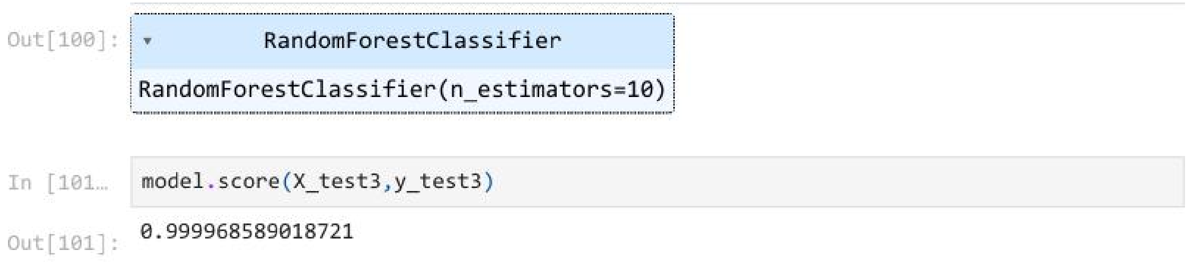
```
In [60]:  LR.score(X_test,y_test)

Out[60]: 0.9654902509380063

In [61]:  LR.predict([[4.8,21.75,78.18,100]])

         C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
         not have valid feature names, but LinearRegression was fitted with feature names
           warnings.warn(
Out[61]: array([94.50592747])

In [62]:  LR.predict([[5.2,7.625,76.53,75.0]])

         C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
         not have valid feature names, but LinearRegression was fitted with feature names
           warnings.warn(
Out[62]: array([68.01686639])

In [64]:  sns.regplot(x=y_test, y=predictions)
          plt.xlabel("True Values")
          plt.ylabel("Predictions")
          plt.title("Regression Plot of Predictions vs True Values")
          plt.show()
```



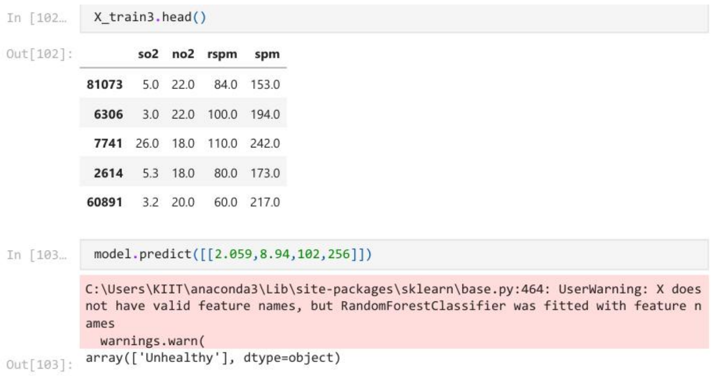
```
In [65]:  print('R^2_Square:%.2f '% r2_score(y_test, predictions))
          print('MSE:%.2f '% np.sqrt(mean_squared_error(y_test, predictions)))

          R^2_Square:0.97
          MSE:11.10

In [66]:  X1= data[['so2','no2','rspm','spm']]
          y1 = data['AQI']
          y.tail()

Out[66]:  99995    237.000000
          99996    148.666667
          99997    168.666667
          99998    178.000000
          99999    203.000000
          Name: AQI, dtype: float64
```



```
Out[94]:  ▾ LogisticRegression
          LogisticRegression()

In [95]:  logmodel2.score(X_test3,y_test3)

Out[95]:  0.9165096117602713

In [96]:  logmodel2.predict([[4.8,17.4,78.48,200]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but LogisticRegression was fitted with feature names
            warnings.warn(
Out[96]:  array(['Poor'], dtype=object)

In [97]:  logmodel2.predict([[67.4,127.7,78.48,215]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but LogisticRegression was fitted with feature names
            warnings.warn(
Out[97]:  array(['Good'], dtype=object)

In [98]:  logmodel2.predict([[2.059,8.94,102,256]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but LogisticRegression was fitted with feature names
            warnings.warn(
Out[98]:  array(['Poor'], dtype=object)
```



```
Out[100]:  ▾          RandomForestClassifier
           RandomForestClassifier(n_estimators=10)

In [101…   model.score(X_test3,y_test3)

Out[101]:  0.999968589018721
```

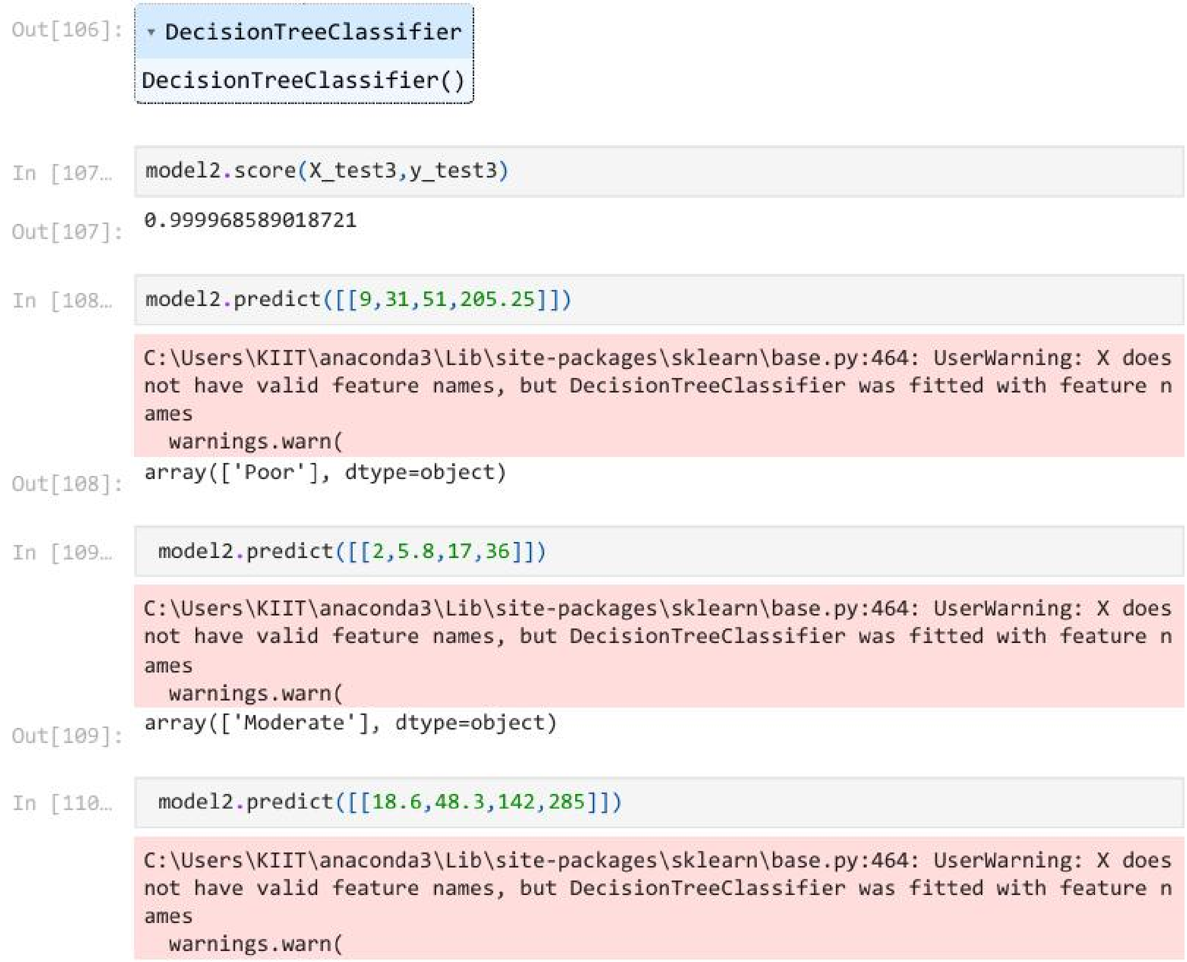


```
In [102…  X_train3.head()
```

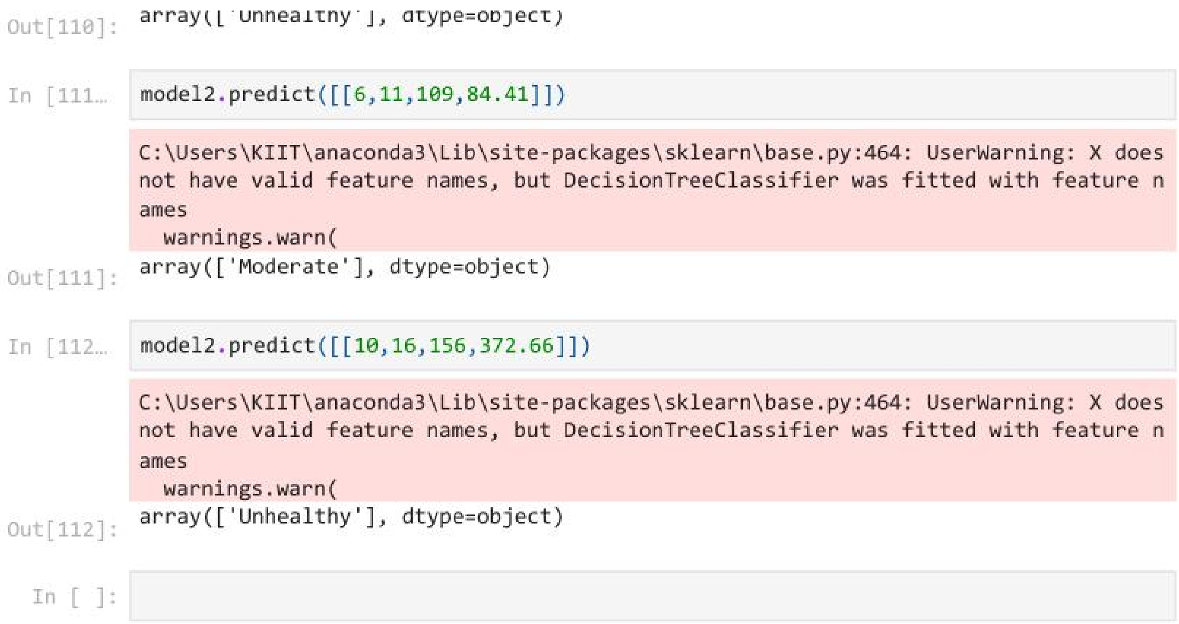
Out[102]:

	so2	no2	rspm	spm
81073	5.0	22.0	84.0	153.0
6306	3.0	22.0	100.0	194.0
7741	26.0	18.0	110.0	242.0
2614	5.3	18.0	80.0	173.0
60891	3.2	20.0	60.0	217.0

```
In [103…  model.predict([[2.059,8.94,102,256]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but RandomForestClassifier was fitted with feature n
          ames
            warnings.warn(
Out[103]: array(['Unhealthy'], dtype=object)
```



```
Out[106]: ▾ DecisionTreeClassifier
          DecisionTreeClassifier()

In [107…  model2.score(X_test3,y_test3)

Out[107]: 0.999968589018721

In [108…  model2.predict([[9,31,51,205.25]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but DecisionTreeClassifier was fitted with feature n
          ames
            warnings.warn(
Out[108]: array(['Poor'], dtype=object)

In [109…   model2.predict([[2,5.8,17,36]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but DecisionTreeClassifier was fitted with feature n
          ames
            warnings.warn(
Out[109]: array(['Moderate'], dtype=object)

In [110…   model2.predict([[18.6,48.3,142,285]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but DecisionTreeClassifier was fitted with feature n
          ames
            warnings.warn(
```



```
Out[110]: array(['Unhealthy'], dtype=object)

In [111…  model2.predict([[6,11,109,84.41]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but DecisionTreeClassifier was fitted with feature n
          ames
            warnings.warn(
Out[111]: array(['Moderate'], dtype=object)

In [112…  model2.predict([[10,16,156,372.66]])

          C:\Users\KIIT\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does
          not have valid feature names, but DecisionTreeClassifier was fitted with feature n
          ames
            warnings.warn(
Out[112]: array(['Unhealthy'], dtype=object)

In [ ]:
```

Chapter 6 Conclusion

The India Air Quality Data Analysis project effectively demonstrates the application of supervised learning algorithms such as Linear Regression, Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. Through this project, the Air Quality Index (AQI) was analyzed and predicted based on key factors, including sulfur dioxide (SO_2), nitrogen dioxide (NO_2), suspended particulate matter (SPM), and respirable suspended particulate matter (RSPM). The project successfully classified air quality levels into categories like good, satisfactory, moderately polluted, poor, very poor, and severe. This work highlights the importance of leveraging machine learning techniques to gain insights into environmental data, paving the way for more informed decisions to improve air quality and public health.

Although this project provides a solid foundation for AQI prediction and classification, there are opportunities for future work. For instance, integrating time-series forecasting models could enable predicting future AQI values based on historical trends, which would be especially valuable for air quality monitoring systems. Additional environmental variables such as temperature, humidity, wind speed, or even traffic data can be integrated to increase the scope of the model.

Chapter 7 References

“Gkatzelis, Georgios I., et al. "The global impacts of COVID-19 lockdowns on urban air pollution: A critical review and recommendations." *Elem Sci Anth* 9.1 (2021): 00176.”

“Gkatzelis, G. I., Gilman, J. B., Brown, S. S., Eskes, H., Gomes, A. R., Lange, A. C., ... & Kiendler-Scharr, A. (2021). The global impacts of COVID-19 lockdowns on urban air pollution: A critical review and recommendations. *Elem Sci Anth*, *9*(1), 00176.”

“Gkatzelis, Georgios I., Jessica B. Gilman, Steven S. Brown, Henk Eskes, A. Rita Gomes, Anne C. Lange, Brian C. McDonald et al. "The global impacts of COVID-19 lockdowns on urban air pollution: A critical review and recommendations." *Elem Sci Anth* 9, no. 1 (2021): 00176.”

“Mahato, Susanta, Swades Pal, and Krishna Gopal Ghosh. "Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India." *Science of the total environment* 730 (2020): 139086.”

“Gupta, N. Srinivasa, et al. "Prediction of air quality index using machine learning techniques: a comparative analysis." *Journal of Environmental and Public Health* 2023.1 (2023): 4916267.”

Chapter 8 Individual Contribution

AIR QUALITY INDEX ANALYSIS

INDRANIL BHATTACHARJEE
21051983

Abstract

This Air Quality Index (AQI) analysis focuses on examining air pollution levels over time and across different locations. My individual contribution involved collecting, processing, and visualizing AQI data to identify patterns, trends, and potential health risks associated with air pollution. I began by gathering raw data from various sources and systematically cleaning and transforming it, ensuring consistency and accuracy. Following this, I calculated AQI values for key pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, and O_3) using established AQI formulas to derive location-specific overall AQI values.

Individual Contribution and findings:

My contribution focused on implementing a decision tree model to predict AQI levels based on pollutants ($PM_{2.5}$, NO_2) and environmental factors (temperature, season). I prepared and tuned the model, achieving high accuracy in classifying AQI categories. Key findings showed that $PM_{2.5}$ and NO_2 were the most influential factors, with air quality worsening in colder months. This model highlights actionable pollutant thresholds and supports proactive air quality management for healthier environments.Also, I have contributed in report making and presentation.

AIR QUALITY INDEX ANALYSIS

SHUBHAM
2105751

Abstract

This Air Quality Index (AQI) analysis focuses on examining air pollution levels over time and across different locations. My individual contribution involved collecting, processing, and visualizing AQI data to identify patterns, trends, and potential health risks associated with air pollution. I began by gathering raw data from various sources and systematically cleaning and transforming it, ensuring consistency and accuracy. Following this, I calculated AQI values for key pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, and O_3) using established AQI formulas to derive location-specific overall AQI values.

Individual Contribution and findings:

My contribution involved developing a linear regression model to predict AQI levels based on pollutant concentrations (e.g., $PM_{2.5}$, NO_2) and environmental factors (temperature, season). I prepared and optimized the model, identifying significant predictors. Key findings showed that $PM_{2.5}$ and NO_2 had the strongest correlation with AQI, and seasonal factors affected overall levels, with AQI worsening in colder months. This model provides insights into pollutant impacts on air quality, supporting targeted interventions.Also,I have contributed in report making and presentation.

AIR QUALITY INDEX ANALYSIS

ROHIT KUMAR
21052984

Abstract

This Air Quality Index (AQI) analysis focuses on examining air pollution levels over time and across different locations. My individual contribution involved collecting, processing, and visualizing AQI data to identify patterns, trends, and potential health risks associated with air pollution. I began by gathering raw data from various sources and systematically cleaning and transforming it, ensuring consistency and accuracy. Following this, I calculated AQI values for key pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, and O_3) using established AQI formulas to derive location-specific overall AQI values.

Individual Contribution and findings:

My contribution centered on implementing a random forest model to predict AQI levels based on pollutant concentrations (e.g., $PM_{2.5}$, NO_2) and environmental factors (season, temperature). I optimized the model and analyzed feature importance. Key findings revealed that $PM_{2.5}$ and NO_2 were top predictors, with AQI worsening in colder months. This model enhances accuracy in AQI prediction and highlights priority pollutants, aiding effective air quality management.Also,I have contributed in report making and presentation.

AIR QUALITY INDEX ANALYSIS

ABHINEET YADAV
21051260

Abstract

This Air Quality Index (AQI) analysis focuses on examining air pollution levels over time and across different locations. My individual contribution involved collecting, processing, and visualizing AQI data to identify patterns, trends, and potential health risks associated with air pollution. I began by gathering raw data from various sources and systematically cleaning and transforming it, ensuring consistency and accuracy. Following this, I calculated AQI values for key pollutants ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, and O_3) using established AQI formulas to derive location-specific overall AQI values.

Individual Contribution and findings:

My contribution involved developing a logistic regression model to classify AQI levels (e.g., Good, Moderate, Unhealthy) based on pollutant concentrations ($PM_{2.5}$, NO_2) and environmental factors (season, temperature). I optimized the model and interpreted the coefficients. Key findings showed $PM_{2.5}$ and seasonality as strong predictors, with higher AQI levels in colder months. This model offers clear insights for categorizing air quality, supporting timely health advisories.Also,I have contributed in report making and presentation.

AIR QUALITY INDEX ANALYSIS

ORIGINALITY REPORT

17%	14%	3%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	2%
2	Submitted to KIIT University Student Paper	2%
3	Submitted to University of Northampton Student Paper	1%
4	github.com Internet Source	1%
5	www.worldleadershipacademy.live Internet Source	1%
6	discovery.researcher.life Internet Source	1%
7	Submitted to University of Liverpool Student Paper	1%
8	Submitted to Midlands State University Student Paper	1%
9	fastercapital.com Internet Source	<1%

10	link.springer.com Internet Source	<1 %
11	www.cameralyze.co Internet Source	<1 %
12	Yizhi Deng, Jie Xu, Bo Zhang, Jinxiang Feng, Jun Gao. "Weber vector local pattern", Optik, 2023 Publication	<1 %
13	Submitted to Polytechnic of Zagreb Student Paper	<1 %
14	dokumen.pub Internet Source	<1 %
15	Submitted to Neath Port-Talbot College Student Paper	<1 %
16	Submitted to University of Sunderland Student Paper	<1 %
17	C.R. Dhivyaa, K. Nithya, T. Vignesh, R. Sudhakar, K. Sathis Kumar, T. Janani. "An Enhanced Deep Learning Model for Tomato Leaf Disease Prediction", 2023 8th International Conference on Communication and Electronics Systems (ICCES), 2023 Publication	<1 %
18	Submitted to University of Greenwich Student Paper	<1 %

19	Submitted to National Institute of Technology Warangal Student Paper	<1 %
20	Submitted to University of Glamorgan Student Paper	<1 %
21	blog.bit.ai Internet Source	<1 %
22	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1 %
23	Submitted to De Montfort University Student Paper	<1 %
24	Submitted to Goldey-Beacom College Student Paper	<1 %
25	Submitted to October University for Modern Sciences and Arts (MSA) Student Paper	<1 %
26	Submitted to The University of Manchester Student Paper	<1 %
27	Submitted to University of Bradford Student Paper	<1 %
28	Submitted to University of West London Student Paper	<1 %
29	cse.kiit.ac.in Internet Source	<1 %

30	Submitted to Kwame Nkrumah University of Science and Technology Student Paper	<1 %
31	Submitted to University of East London Student Paper	<1 %
32	Submitted to University of Winchester Student Paper	<1 %
33	www.lambdatest.com Internet Source	<1 %
34	Submitted to University of Technology Bahrain Student Paper	<1 %
35	builtin.com Internet Source	<1 %
36	repairit.wondershare.com Internet Source	<1 %
37	www.mdpi.com Internet Source	<1 %
38	Submitted to University of Essex Student Paper	<1 %
39	irep.iium.edu.my Internet Source	<1 %
40	www.ijeat.org Internet Source	<1 %

41	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
42	repo.lib.tut.ac.jp Internet Source	<1 %
43	www.mygreatlearning.com Internet Source	<1 %
44	www.slideshare.net Internet Source	<1 %
45	www.zenatix.com Internet Source	<1 %
46	affiliatepal.net Internet Source	<1 %
47	core.ac.uk Internet Source	<1 %
48	www.researchgate.net Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		