

# AIR QUALITY PREDICTION USING MACHINE LEARNING ALGORITHMS

PONNAGANTI SIRISHA<sup>1</sup>, VEMULA B S SAI RAJA<sup>2</sup>, KOMMURI JAGANNADH<sup>3</sup>

<sup>1</sup>Ponnaganti Sirisha, Department of Information Technology & Dhanekula Institute of Engineering and Technology

<sup>2</sup>Vemula B S Sai Raja, Department of Information Technology & Dhanekula Institute of Engineering and Technology

<sup>3</sup>Kommuri Jagannadh, Department of Information Technology & Dhanekula Institute of Engineering and Technology

\*\*\*

**Abstract** - Days gone by air pollution is rapidly increasing, pollution occurs due to human activities, industrialization and burning of fossil fuels. The air is polluted by dangerous gases present in the atmosphere; the pollutants are carbon dioxide (CO<sub>2</sub>), carbon monoxide (CO), Sulphur dioxide (SO<sub>2</sub>), nitrogen dioxide (NO<sub>2</sub>), etc. The pollutants of which Particulate Matter (PM 2.5) consists of suspended particles with a diameter of less than 2.5 micrometers are considered harmful. Air pollution is a major issue that impacts humans and causes health problems including heart disease, lung cancer, and respiratory diseases such as emphysema and asthma. previous techniques such as probability, statistics, traditional methods, etc. These methods are too complex to predict. Machine learning (ML) is the best way to predict air quality. In this paper, we need to implement models that record information about air pollutant concentrations. We compare machine learning algorithms like linear regression (LR), random forest (RF), k-nearest neighbor (KNN), and decision tree regression (DT) to predict the air quality index (AQI).

**Key Words:** Machine learning- Linear Regression, Random Forest, k-nearest neighbor, and Decision Tree regression

## 1.INTRODUCTION

Air pollution a serious environmental problem that contributes to global warming has a greater impact on human health that causing premature death from cancer, respiratory disease, or heart disease. Poor air quality contributes to modern environmental problems such as global warming, acid rain, reduced visibility, climate change, etc. The Air Quality Index (AQI) is a measure that describes the air quality levels based on the concentrations of several pollutants in the atmosphere, usually PM<sub>2.5</sub>, PM<sub>10</sub>, carbon monoxide (co), Sulphur dioxide (so<sub>2</sub>), nitrogen dioxide (no<sub>2</sub>), and ozone (o<sub>3</sub>).

Air pollution is caused by two types of pollutants: primary pollutants and secondary pollutants. The primary pollutants include carbon dioxide (CO<sub>2</sub>), Sulphur oxide (SOX), nitrogen oxide (NOX), carbon monoxide (CO), and chlorofluorocarbons (CFC).and the secondary pollutants include ground-level ozone and acid rain.

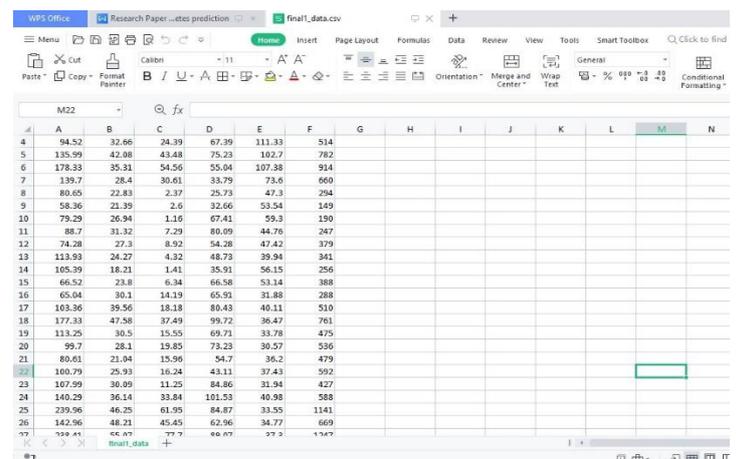
We have collected datasets on total pollution in different regions of India. To begin, we compute the individual pollution index. For each marked card and find the corresponding AQI for the area. We have provided a model to predict the air quality index and our model can predict the air quality in any region of India.

By predicting the air quality index, we can identify major pollution bottlenecks and pollution affected areas across India. In this forecasting model various data are extracted using various techniques to find out the most affected areas in the country.

## 2. METHODOLOGY

### 2.1 DATASET

In this dataset we have used parameters like particulate matter, nitrogen dioxide, carbon monoxide, Sulphur dioxide, Ozone



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
4	94.52	32.96	24.39	67.39	111.33	514								
5	135.99	42.88	43.48	75.23	102.7	782								
6	178.33	35.31	54.56	55.04	107.38	914								
7	139.7	28.4	30.61	33.79	73.6	660								
8	80.65	22.83	2.37	25.73	47.3	294								
9	58.36	21.39	2.6	32.66	53.54	149								
10	79.29	25.94	1.16	67.41	59.3	150								
11	88.7	31.32	7.29	80.09	44.76	247								
12	74.28	27.3	8.92	54.28	47.42	379								
13	113.99	24.27	4.32	48.73	39.94	341								
14	165.99	18.21	1.41	35.91	56.15	256								
15	66.52	23.8	6.34	66.58	53.14	388								
16	65.04	30.1	14.19	65.91	31.88	288								
17	103.36	39.56	18.18	80.43	40.11	510								
18	177.33	47.58	37.49	99.72	36.47	761								
19	113.25	30.5	15.55	69.71	33.78	475								
20	89.7	28.1	19.85	73.23	30.57	536								
21	89.61	21.04	15.96	54.7	36.2	479								
22	160.79	25.93	16.24	43.11	37.43	592								
23	107.99	36.09	11.25	84.86	31.94	427								
24	140.29	36.14	33.84	101.53	46.96	588								
25	239.96	46.25	61.95	84.87	33.55	1141								
26	142.96	48.21	45.45	62.96	34.77	669								
27	120.41	66.87	77.7	88.87	37.3	1147								

### 2.2 DATA-PREPROCESSING

#### Step:1

In this step I have completed the preprocessing of the given dataset by using the pandas. By using the pandas, we loaded the dataset into the Jupiter notebook.

```
data=pd.read_csv("file path")
```

#### Step:2

Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore, this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.

```
data.isna().sum()
PM2.5    0
NO2      0
CO        0
SO2       0
O3        0
AQI       0
dtype: int64
```

Checking whether the dataset is having null values are not

**Step:3** Apply some statistical functions on the dataset

```
data.describe()
count    22618.000000    22618.000000    22618.000000    22618.000000    22618.000000    22618.000000
mean     67.756028      29.718371      2.341847      13.891994      35.123085      167.385047
std      63.404533      24.503292      6.965907      16.661959      21.604953      140.384503
min       0.160000      0.010000      0.000000      0.010000      0.010000      14.000000
25%      29.032500      12.880000      0.610000      5.830000      19.560000      81.000000
50%      48.855000      23.220000      0.940000      9.220000      31.625000      118.000000
75%      81.340000      39.030000      1.470000      14.707500      46.250000      211.000000
max     914.940000     362.210000     175.810000     186.080000     257.730000     1917.000000
```

data.describe().

**Step:4** Assign the values to x and y:

Because x is the independent variable and the y is the dependent variable.

x	PM2.5	NO2	CO	SO2	O3
0	83.13	28.71	6.93	49.52	59.76
1	79.84	28.68	13.85	48.49	97.07
2	94.52	32.66	24.39	67.39	111.33
3	135.99	42.08	43.48	75.23	102.70
4	178.33	35.31	54.56	55.04	107.38
...	...	...	...	...	...
22613	15.02	25.06	0.47	8.55	23.30
22614	24.38	26.06	0.52	12.72	30.14
22615	22.91	29.53	0.48	8.42	30.96
22616	16.64	29.26	0.52	9.84	28.30
22617	15.00	26.85	0.59	2.10	17.05

**X values**

y	
0	209.0
1	328.0
2	514.0
3	782.0
4	914.0
...	...
22613	41.0
22614	70.0
22615	68.0
22616	54.0
22617	50.0

Name: AQI, Length: 22618, dtype: float64

**Y values**

**Step:5** Split the data for testing and training:

```
#Splitting Data
X = data.iloc[:, :-1] #Independent features
y = data.iloc[:, -1] #Dependent feature

#Train Test Splitting
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

**2.3 ALGORITHMS**

The algorithms we used – Linear Regression, Decision Tree Regression, Random Forest, and k-nearest Neighbors algorithm

**Linear Regression**

Linear regression assumes that there is a linear relationship between the independent variables and the dependent variable, and that the independent variables are not highly correlated with each other.

**Decision Tree Regression**

Decision tree is a machine learning algorithm that can be used for both classification and regression tasks. The algorithm creates a tree-like model of decisions and their possible consequences.

**Random Forest**

Random forest is a machine learning algorithm that is used for classification, regression, and other tasks. It is a type of ensemble learning method, which combines multiple decision trees to improve the accuracy of predictions.

**k-nearest Neighbors algorithm**

KNN (k-Nearest Neighbors) is a simple, non-parametric algorithm used for classification and regression. The algorithm works by finding the k nearest data points to the test point and then predicting the label or value of the test point based on the labels or values of its nearest neighbors

**DERIVATIONS**

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}$$

$$= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

MSE = mean squared error  
n = number of data points  
Y<sub>i</sub> = observed values  
Ŷ<sub>i</sub> = predicted values

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

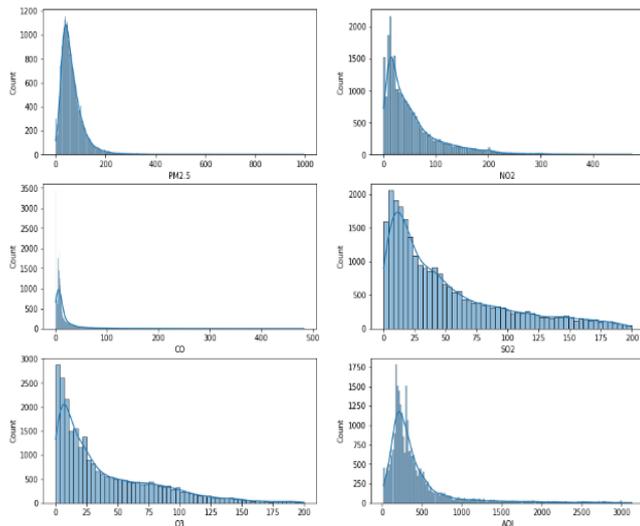
MAE = mean absolute error

$y_i$  = prediction

$x_i$  = true value

$n$  = total number of data points

### 2.4 VISUALIZATION

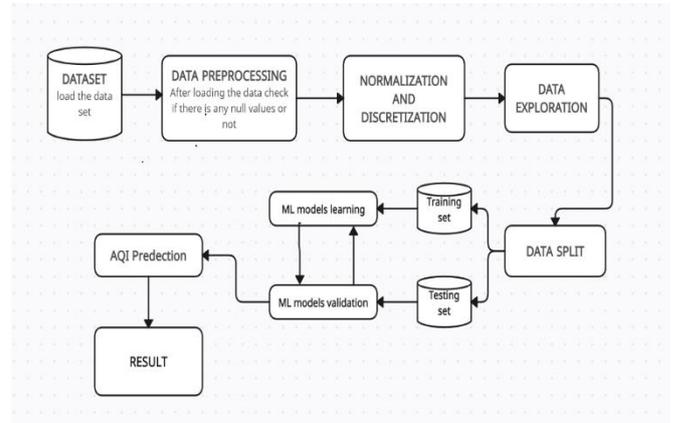


The flow of pollutants percentage in the atmosphere are show in the form of graphical representation

### 3.PROPOSED SYSTEM

The Air pollutants information is retrieved using Web Scrapping and extract the PM2.5 and then stored as a dataset. This dataset is preprocessed with various functions such as normalization, attribute selection and discretization. Once the dataset is created it is divided into training datasets and test datasets. and more supervised machine learning algorithms applied to training datasets. The obtained results were compared with the test data set and the results were analyzed. For air quality index prediction using Supervised Machine Learning approach we consider Regression techniques such as Linear Regression, k-nearest neighbor, Decision tree Regression and Random-forest Regression to predict the AQI.

### 3.1 PROCEDURE



### 3.2 EQUVALUES

AIR QUALITY INDEX (AQI)	CATEGORY
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

### 4.CONCLUSION AND RESULTS

#### 4.1 RESULTS

Decision Tree: r2\_score

```
r2_score(y_test, prediction)
```

0.778979238122161

Decision Tree-Part:2 after hyper parameter tuning

```
# Fitting Model without any tuning
model = RandomForestRegressor(n_estimators = 200, random_state = 0)
model = model.fit(X_train, y_train)
prediction = model.predict(X_test)

print("Coefficient of Determination (R^2) for train dataset: ", model.score(X_train, y_train))
print("Coefficient of Determination (R^2) for test dataset: ", model.score(X_test, y_test))

print('MAE:', metrics.mean_absolute_error(y_test, prediction))
print('MSE:', metrics.mean_squared_error(y_test, prediction))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, prediction)))
```

Coefficient of Determination (R<sup>2</sup>) for train dataset: 0.9878910757554696  
 Coefficient of Determination (R<sup>2</sup>) for test dataset: 0.9165444529936397  
 MAE: 20.956891577477897  
 MSE: 1067.226112075384  
 RMSE: 32.6684268380861

### k-nearest Neighbors algorithm

```
from sklearn import metrics
from sklearn.metrics import r2_score
print('r2_score:', r2_score(ytest, ypred_knn))
r2_score: 0.5539583898980076
```

### Random Forest

```
print("Coefficient of Determination (R^2) for train dataset: ", best_rf.score(X_train, y_train))
print("Coefficient of Determination (R^2) for test dataset: ", best_rf.score(X_test, y_test))
```

Coefficient of Determination (R<sup>2</sup>) for train dataset: 0.9198317555000328  
Coefficient of Determination (R<sup>2</sup>) for test dataset: 0.8939193526213225

### R2 Score:

	Algorithm	R <sup>2</sup> Score
0	Linear Regression	0.825624
1	KNN	0.851665
2	Random Forest	0.876816
3	Decision Tree	0.806792

## 5. CONCLUSIONS

In this paper we proposed a model based on the ML algorithm like Linear regressor, Decision tree regressor, Random Forest, K-Nearest neighbor which was used to predict the quality of air in the atmosphere, the model take the data as input from dataset and perform the preprocessing and calculate the quality of air, this model is efficient and gave results with high performance.

## 6. ACKNOWLEDGMENT

For allowing the authors to perform this project AIR QUALITY PREDECTION USING MACHINE LEARNING ALGORITHMS by Dhanekula Institute of Engineering & Technology, Ganguru, Bachelor of Technology, Faculty of Information Technology Department, and our beloved guide Associate Professor Dr. K. Sandeep through the provision of computational resources and a conducive working environment.

## REFERENCES:

- [1]. Kostandina Veljanovska<sup>1</sup> & Angel Dimoski<sup>2</sup>, Air Quality Index Prediction Using Simple Machine Learning Algorithms, 2018, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).
- [2]. Xiaosong Zhao, Rui Zhang, Jheng-Long Wu, Pei-Chann Chang and Yuan ZeUniversity, A Deep Recurrent Neural Network for Air Quality Classification, 2018, Journal of Information Hiding and Multimedia Signal Processing
- [3]. Savita Vivek Mohurle, Dr. Richa Purohit and Manisha Patil, A study of fuzzy clustering concept for measuring air pollution index, 2018, International Journal of Advanced Science and Research
- [4]. Aditya C R, Chandana R Deshmukh, Nayana D K and Praveen Gandhi Vidyavastu, Detection and Prediction of Air Pollution using Machine Learning Models, 2018, International Journal of Engineering Trends and Technology (IJETT)