PREDICTION OF AIR POLLUTION USING MACHINE LEARNING

[1]**Harshi Pogadadanda**
Final Year B.Tech, Dept. of CSE
SRM Institute of Science and Technology
Email: harshipogadadanda@gmail.com
[2]**Mr. M. Arul Prakash**
Assistant Professor, Dept. of CSE, Kattankulathur Campus
SRM Institute of Science and Technology
Email: arulpram@srmist.edu.in

ABSTRACT

Generally, Air pollution alludes to the issue of toxins into the air that are harmful to human and the entire planet. It can be described as one of the most dangerous threats that the humanity ever faced. It causes damage to animals, crops, forests etc. To prevent this problem in transport sectors have to predict air quality from pollutants using machine learning techniques. Subsequently, air quality assessment and prediction has turned into a significant research zone. The aim is to investigate machine learning based techniques for air quality prediction. In the populated and developing countries, governments consider the regulation of air as a major task. The meteorological and traffic factors, burning of fossil fuels, industrial parameters such as power plant emissions play significant roles in air pollution. Among all the particulate matter that determine the quality of the air, Particulate matter (PM 2.5) needs more attention. When it's level is high in the air, it causes serious issues on people's health. Hence, controlling it by constantly keeping a check on its level in the air is important. In this Linear regression is employed to predict future values of PM2.5 based on the previous PM2.5readings. Knowledge of level of PM2.5 in nearing years, month or week, enables us to reduce its level to lesser than the harmful range. This system attempts to predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city.

I. INTRODUCTION

Particulate matter can be either human-made or naturally occur. Some examples include dust, ash and sea-spray. Particulate matter (including soot) is emitted during the combustion of solid and liquid fuels, such as for power generation, domestic heating and in vehicle engines. Particulate matter varies in size (i.e. the diameter or width of the particle). PM2.5 refers to the mass per cubic meter of air of particles with a size (diameter) generally less than 2.5 micrometers (μm). PM2.5 is also known as fine particulate matter (2.5 micrometers is one 400th of a millimeter). Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. Different machine learning models have been applied to detect air pollution and predict PM2.5 levels based on a data set consisting of daily atmospheric conditions. With economic development and population rise in cities, environmental pollution problems involving air pollution, water pollution, noise and the shortage of land resources have attracted increasing attention. Among these, air pollution's direct impact on human health through exposure to pollutants has resulted in an increased public awareness in both developing and developed countries. Air pollution is usually caused by energy production from power plants, industries, residential heating, fuel burning vehicles, natural disasters, etc. Increasing amounts of potentially harmful gases and particulates are being emitted into the atmosphere on a global scale resulting in damages to human health and the environment. Human health concern is one of the important consequences of air pollution, especially in urban areas. The global warming from anthropogenic greenhouse gas emissions is a long term consequence of air pollution. Accurate air quality forecasting can reduce the effect of a pollution peak on the surrounding population and ecosystem, hence improving air quality forecasting is an important goal for society.

II. BACKGROUND WORK

Air pollution is the introduction of particulates, biological molecules, or other harmful materials into the Earth's atmosphere, causing disease, death to humans, damage to other living organisms such as food crops, or damage to the natural or man-made environment. An air pollutant is a substance in the air that can have adverse effects on humans and the ecosystem. The substance can be solid particles, liquid droplets, or gases. Pollutants are classified as primary or secondary. Primary pollutants are usually produced from a process, such as ash from a volcanic eruption. Other examples include carbon monoxide gas from motor vehicle exhaust, or sulfur dioxide released from factories. Secondary pollutants are not emitted directly. Rather, they form in the air when primary pollutants react or interact. Ground level ozone is a prominent example of a secondary pollutant. The six "criteria pollutants" are ground level ozone (O3), fine particulate matter (PM2.5), carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), and lead, among which ground level O3, PM2.5 and NO2 (main component of NOx) are the most widespread health threats. Ground level O3, a gaseous secondary air pollutant formed by complex chemical reactions between NOx and volatile organic compounds (VOCs) in the

 ISSN: 2278-4632
Vol-10 Issue-10 No.02 October 2020

atmosphere, can have significant negative impacts on human health. Prolonged exposure to O3 concentrations over a certain level may cause permanent lung damage, aggravated asthma, or other respiratory illnesses. Ground level O3 can also have detrimental effects on plants and ecosystems, including damage to plants, reductions of crop yield and increase of vegetation vulnerability to disease.

Particle pollution (also called particulate matter or PM) is the term for a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Others are so small they can only be detected using an electron microscope. Fine particulate matter (PM2.5) consisting of particles with diameter 2.5 μm or smaller, is an important pollutant among the criteria pollutants. The microscopic particles in PM2.5 can penetrate deeply into the lungs and cause health problems, including the decrease of lung function, development of chronic bronchitis and nonfatal heart attacks. Fine particles can be carried over long distances by wind and then deposited on ground or water through dry or wet deposition. The wet deposition is often acidic, as fine particles containing sulfuric acid contribute to rain acidity, or acid rain. The effects of acid rain include changing the nutrient balance in water and soil, damaging sensitive forests and farm crops, and affecting the diversity of ecosystems. PM2.5 pollution is also the main cause of reduced visibility.

The ambient air in most large Indian cities is severely polluted and this pollution has a tremendous impact not only on the health of the population but also in the ecosystem. Industrialization, the growth in the number of vehicles in urban areas has led to a rapid determination of ambient air quality by emitting various kinds of air pollutants. Urban air pollution has grown in cities like Delhi, Mumbai, and Kolkata, across the Indian subcontinent in the last decade in an alarming condition. The World Health Organization ranked Delhi as the fourth-most polluted mega city of the world. However, in Indian subcontinent, it is not just Delhi, but even small and medium towns are deteriorating air quality rapidly. Out of the 23 mega cities, Delhi is the most polluted followed by Mumbai, Calcutta, Bangalore, Chennai, Kanpur, Ahmedabad and Nagpur. They have severe air pollution problems mainly with the average levels of suspended particulate matter levels much higher than the prescribed standards.

Main Objective: The primary goal is to predict air pollution level in City with the ground data set.

- Detects the levels of PM2.5 based on given atmospheric values.
- Predicts the level of PM2.5 for a particular date.

ISSN: 2278-4632
Vol-10 Issue-10 No.02 October 2020

III. PROPOSED APPROACH

The proposed system predicts the pm2.5 level based on a dataset consisting of atmospheric conditions. The system predicts the pm2.5 level for a particular date. A systematic approach has been followed in this analysis which is depicted in figure 3. The approach starts with the collection of dataset from kaggle. Collected data has been preprocessed to remove the redundancy. Preprocessing of data includes steps like parsing of dates, noise removal, cleaning, training and scaling. Data visualization visualizes the data and then apply regression algorithm and finally forecasting pm2.5 value.

MACHINE-LEARNING APPROACH

DATASET

Dataset/Source: Kaggle Structured/Unstructured data:Structured Data in CSV format. Dataset **Description:** The dataset consists of around 450000 records of all the states of India.

The data has been collected from kaggle. The dataset contains twelve attributes: year, month, day, hour, dew point, temperature, pressure, iws, pm 2.5 and predicted pm2.5. The 'Date' describes the sampling date and other parameters give their individual concentration in air.

A	B	C	D	E	F	G	H	I	J
2010	1	2	0	-16	-4	1020	1.79	129	148
2010	1	2	1	-15	-4	1020	2.68	148	159
2010	1	2	2	-11	-5	1021	3.57	159	181
2010	1	2	3	-7	-5	1022	5.36	181	138
2010	1	2	4	-7	-5	1022	6.25	138	109
2010	1	2	5	-7	-6	1022	7.14	109	105
2010	1	2	6	-7	-6	1023	8.93	105	124
2010	1	2	7	-7	-5	1024	10.72	124	120
2010	1	2	8	-8	-6	1024	12.51	120	132
2010	1	2	9	-7	-5	1025	14.3	132	140
2010	1	2	10	-7	-5	1026	17.43	140	152
2010	1	2	11	-8	-5	1026	20.56	152	148
2010	1	2	12	-8	-5	1026	23.69	148	164
2010	1	2	13	-8	-5	1025	27.71	164	158
2010	1	2	14	-9	-5	1025	31.73	158	154
2010	1	2	15	-9	-5	1025	35.75	154	159
2010	1	2	16	-9	-5	1026	37.54	159	164
2010	1	2	17	-8	-5	1027	39.33	164	170
2010	1	2	18	-8	-5	1027	42.46	170	149
2010	1	2	19	-8	-5	1028	44.25	149	154
2010	1	2	20	-7	-5	1028	46.04	154	164
2010	1	2	21	-7	-5	1027	49.17	164	156
2010	1	2	22	-8	-6	1028	52.3	156	126

Figure of sample values of the parameters

Station code is a code given to each station that recorded the data, sampling date is the date when the data is recorded state and location represents state and cities whose data is recorded and agency is the name of agency that recorded the data. Type states the type of area where the data was recorded such as industrial,residential,etc.so2,no2,rspm and spm is the amount of sulphur dioxide, nitrogen dioxide, respirable suspended particulate matter and suspended particulate matter measured respectively.date is a cleaner version of sampling_date. PM2.5 refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair.But majority of values in this column are null.

Splitting for Testing: Data Splitting was done as 80% for training and 20% for testing.

Preprocessing and Feature Selection

We only studied and applied algorithms on the data of Maharashtra State .Hence, no. of rows was reduced to 60,383 and state column automatically is of no more use. All the values in pm2_5 were null values, so we dropped the column. The agency's name have nothing to do with how much polluted the state is. Similarly, stn_code is also not useful. The date is a cleaner representation of sampling_date attribute and so we will eliminate the redundancy by removing the latter. location_monitoring_station attribute is again unnecessary as it contains the location of the monitoring station which we do not need to consider for the analysis.

So, to summarize we have deleted the following features from our dataset: state,pm2_5,agency, stn_code, sampling_date and location_monitoring_station We have simplified the type attribute to contain only one of the three categories: industrial, residential, other. For SO2 and NO2, we replaced nan values by mean. For date, we have dropped nan values as there were only 3 null values.

Data visualization In this step data is visualize by different charts and graphs.

Methodology There are two primary phases in the system:

1. Training phase: The system is trained by using the data in the data set and fits a model based on the algorithm chosen accordingly.
2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked

There are many approaches that can be used for machine learning and data analytics. In this paper we compare three mainstream approaches: linear regression analytics, artificial neural network analytics and long, short term memory (LSTM) analytics.

LINEAR REGRESSION ANALYTICS

Linear regression is a classic approach to model the relationship between a variable and a scalar, which corresponds to features and results in a given data set. The generic equation for linear regression is given as:

$$y=X\text{þ}+s$$

In the above equation, y is the target value, X is an input variable which can be a variable or a matrix. þ and s are matrix weights and associated bias respectively. Both of these are trainable variables. There are many estimation methods to calculate þ and s. The most widespread one is Least-Squares estimation, which minimizes the sum of squared residuals. The equation is shown as:

$$\text{þ}^\wedge = (XT\ X)–1XT\ y$$

Different methods can be used to enhance the performance in identifying þ and s. In this project, the linear regression was implemented with the curve fit tool of Matlab. Matlab uses gradient descent as an estimation method.

A simple linear regression is a basic linear regression where the degree of features is one. The coefficients of a simple linear regression here are [0.02494,6.383] and hence the linear model is given by f(x)=0.02494*x+6.383. The root mean squared error (RMSE) is 6.0882, which provides a better result compared to various other papers in this domain [22-25]. This result is determined by the distribution of samples. It is noted that some outliers exist, e.g. they are extremely high and do not show strong relevance to traffic volume.

To fit this model more accurately, polynomial regression models with degree 2 to degree 6 were considered. A model with degree 6 is the best with RMSE 6.0832 and coefficients [2.391e-08, 3.801e-06, 0.0002249,-0.006186, 0.0806, -0.4151, 6.906] was used for themodel.

Decision Tree Regressor

It is the decision tree regressor function used to build a decision tree model in Machine Learning using Python. The DecisionTreeRegressor () function looks like this:DecisionTreeRegressor (criterion = ‘mse’, random_state =None ,max_depth=None, min_samples_leaf=1,)

- **Criterion:** This function is used to measure the quality of a split in the decision tree regression. By default, it is 'mse' (the mean squared error), and it also supports 'mae' (the mean absoluteerror).
- **max_depth:** This is used to add maximum depth to the decision tree after the tree is expanded.
- **min_samples_leaf:** This function is used to add the minimum number of samples required to be present at a leafnode.

IV. MAIN CODE

```
from sklearn import preprocessing
import numpy as np
import pandas
import matplotlib.pyplot as plt
from   sklearn   import   model_selection
from sklearn import linear_model
from   sklearn.metrics   import   classification_report
from sklearn.metrics import accuracy_score
import seaborn as sns
url = "/content/drive/My Drive/prsa.csv"
names   =   ['year','month','day','hour','DEWP','TEMP','PRES','Iws','pm2.5','Predicted   pm2.5']
dataset = pandas.read_csv(url,names=names)
dataset.head()
dataset.info()
df_train_labels     =     dataset[['TEMP','PRES','pm2.5']]
columnsList_num = ['TEMP', 'PRES', 'pm2.5']
for i in columnsList_num: var = i
   data = pandas.concat([dataset['pm2.5'], dataset[var]], axis=1)
   data.plot.scatter(x=var, y='pm2.5', ylim=(dataset['pm2.5'].min(),dataset['pm2.5'].max()))
from   scipy.stats   import   norm
from scipy import stats
sns.distplot(np.log(dataset['PRES']), fit=norm) fig = plt.figure()
res = stats.probplot((dataset['PRES']), plot=plt)
# create a figure and axis fig,
ax = plt.subplots()
```

scatter the sepal_length against the sepal_width

ax.scatter(dataset['year'], dataset['Predicted pm2.5']) #

set a title and labels

ax.set_title('pollution dataset predicted values')

ax.set_xlabel('year')

ax.set_ylabel('Predicted pm2.5') fig, ax = plt.subplots()

ax.scatter(Y_validation, y_pred, edgecolors=(0, 0, 0))

ax.set_xlabel('Measured')

ax.set_ylabel('Predicted')

V. RESULTS

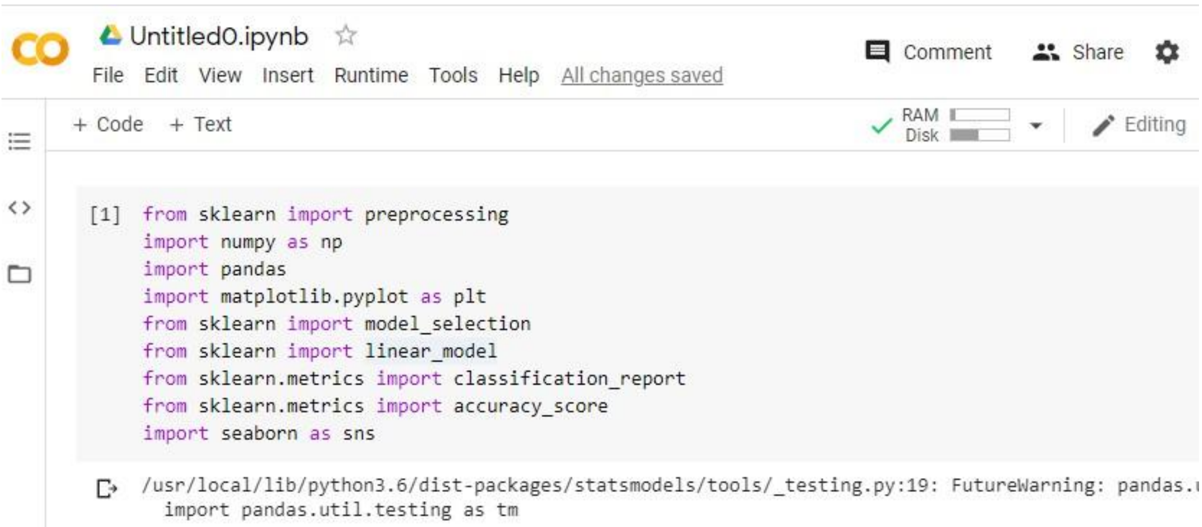


Fig: Importing packages

Uploading dataset

```
url = "/content/drive/My Drive/prsa.csv"
names = ['year','month','day','hour','DEWP','TEMP','PRES','Iws','pm2.5','Predicted pm2.5
dataset = pandas.read_csv(url,names=names)
dataset.head()
```

	year	month	day	hour	DEWP	TEMP	PRES	Iws	pm2.5	Predicted pm2.5
0	2010	1	2	0	-16	-4.0	1020.0	1.79	129	148
1	2010	1	2	1	-15	-4.0	1020.0	2.68	148	159
2	2010	1	2	2	-11	-5.0	1021.0	3.57	159	181
3	2010	1	2	3	-7	-5.0	1022.0	5.36	181	138
4	2010	1	2	4	-7	-5.0	1022.0	6.25	138	109

Fig: Uploading dataset

Dataset information

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41547 entries, 0 to 41546
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   year            41547 non-null  int64
 1   month           41547 non-null  int64
 2   day             41547 non-null  int64
 3   hour            41547 non-null  int64
 4   DEWP            41547 non-null  int64
 5   TEMP            41547 non-null  float64
 6   PRES            41547 non-null  float64
 7   Iws             41547 non-null  float64
 8   pm2.5           41547 non-null  int64
 9   Predicted pm2.5 41547 non-null  int64
dtypes: float64(3), int64(7)
memory usage: 3.2 MB
```

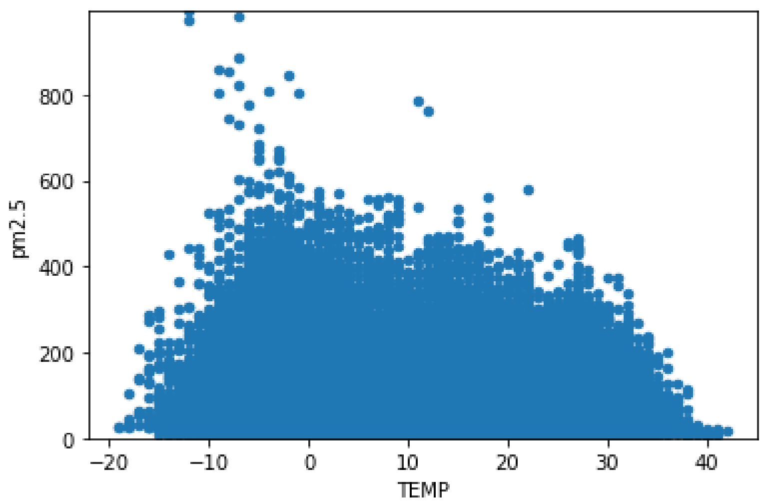
Fig: Dataset information

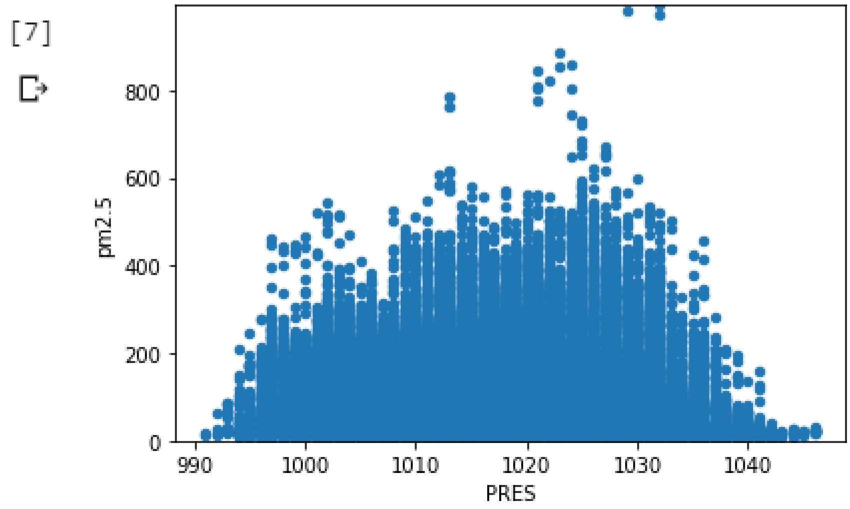
Visualizing the data

```
[5] df_train_labels = dataset[['TEMP','PRES','pm2.5']]
```

```
[7] columnsList_num = ['TEMP', 'PRES', 'pm2.5']
    for i in columnsList_num:
        var = i
        data = pandas.concat([dataset['pm2.5'], dataset[var]], axis=1)
        data.plot.scatter(x=var, y='pm2.5', ylim=(dataset['pm2.5'].min(),dataset['pm2.5'].max()))
```

[7]





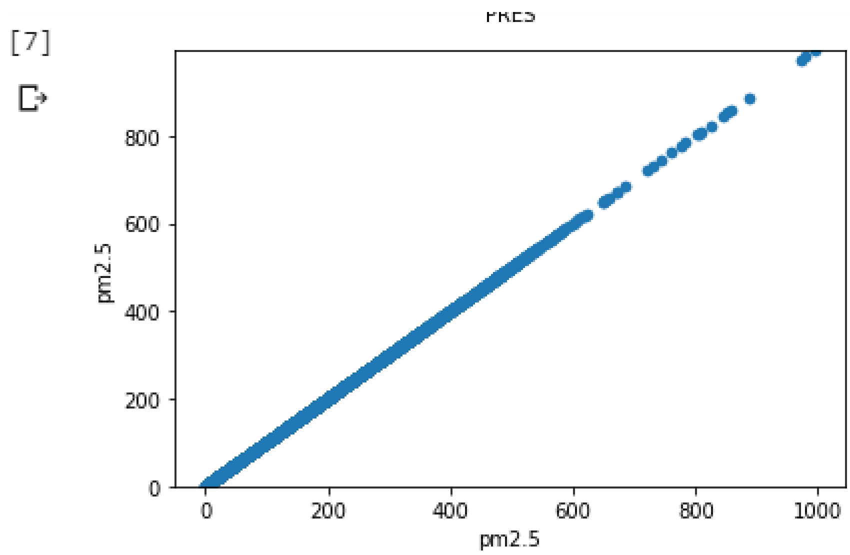
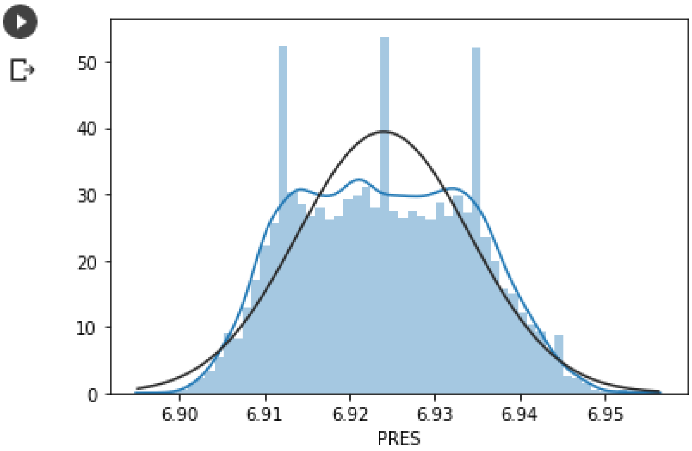
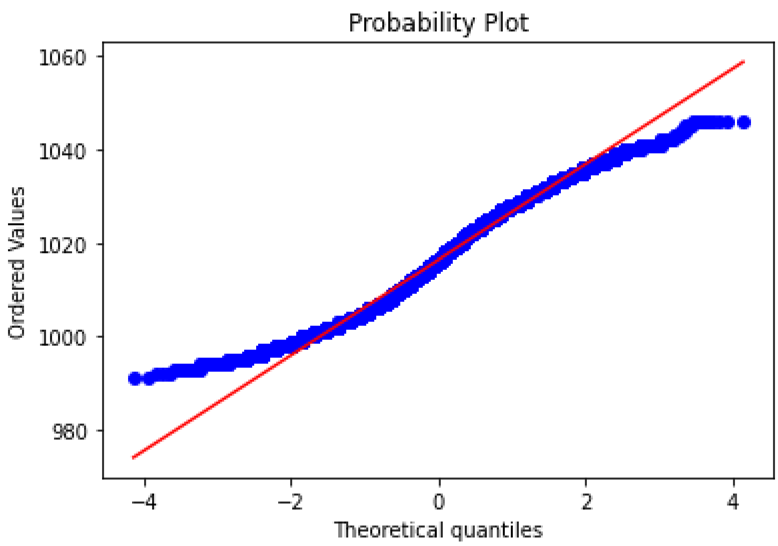


Fig: Visualizing data

```
from scipy.stats import norm
from scipy import stats
sns.distplot(np.log(dataset['PRES']), fit=norm)
fig = plt.figure()
res = stats.probplot((dataset['PRES']), plot=plt)
```

ISSN: 2278-4632
Vol-10 Issue-10 No.02 October 2020

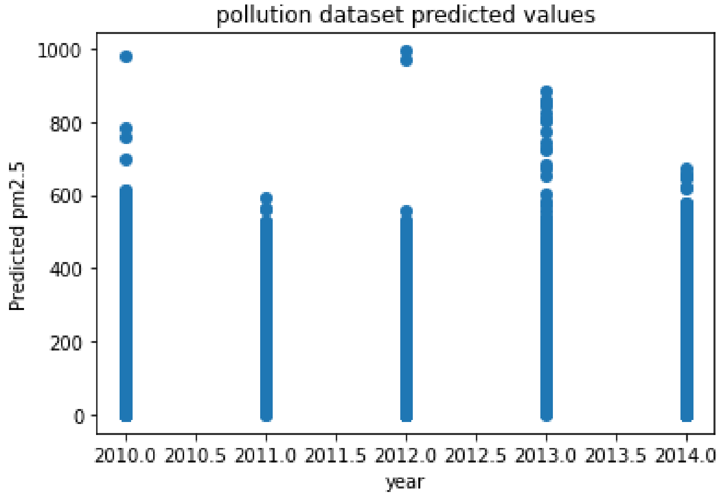





```
[9] # create a figure and axis
    fig, ax = plt.subplots()

    # scatter the sepal_length against the sepal_width
    ax.scatter(dataset['year'], dataset['Predicted pm2.5'])
    # set a title and labels
    ax.set_title('pollution dataset predicted values')
    ax.set_xlabel('year')
    ax.set_ylabel('Predicted pm2.5')
```

```
[9] Text(0, 0.5, 'Predicted pm2.5')
```




```
fig, ax = plt.subplots()
ax.scatter(Y_validation, y_pred, edgecolors=(0, 0, 0))
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```

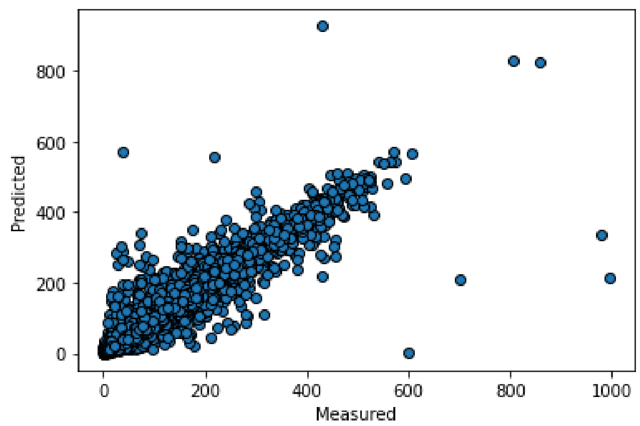


Fig: Visualizing data

ISSN: 2278-4632
Vol-10 Issue-10 No.02 October 2020

Predicting value based on dataset:

```
[11] example_measures = np.array([2050,4,14,21,-7,-5,1090,1.17,20])
     example_measures=example_measures.reshape(1 ,-1)
```

```
[12] reg=linear_model.LinearRegression()
     reg.fit(X_train,Y_train)
     reg.intercept_
     print('variance_score: %.2f' % reg.score(X_validation,Y_validation))
     y_pred=reg.predict(X_validation)
     predic = reg.predict(example_measures)
     print(predic)
```

```
variance_score: 0.92
[31.92260954]
```

Checking prediction value is harmful or not and r2 score for linear regression:

```
[13] if(predic>=45):
         print("harmful")
     else:
         print("not harmful")
```

```
not harmful
```

```
[ ] print("R^2 score for liner regression: ", reg.score(X_validation, Y_validation))
```

```
R^2 score for liner regression:  0.9174293611218931
```

Decision tree regression and r2 score:

```
[ ] from sklearn.tree import DecisionTreeRegressor
    dtr = DecisionTreeRegressor()
    dtr.fit(X_train, Y_train)
```

```
DecisionTreeRegressor(ccp_alpha=0.0, criterion='mse', max_depth=None,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort='deprecated',
                      random_state=None, splitter='best')
```

```
print("Coefficient of determination R^2 <-- on test set: {}".format(dtr.score(X_validation, Y_validation)))
```

```
Coefficient of determination R^2 <-- on test set: 0.8532710701705802
```

VI. CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The results show that machine learning models can be efficiently used to predict the level of PM2.5 in the future. The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities.

VII. REFERENCES

i. http://ijettjournal.org/2018/volume-59/number-4/IJETT-V59P238.pdf

ii. Pandey, Gaurav, Bin Zhang, and Le Jian. " Predicting sub-micron air pollution indicators: a machine learning approach." ; Environmental Science: Processes & amp; Impacts15.5

iii. Dan wei: Predicting air pollution level in a specific city[2014]

iv. Dixian Zhu, ChangjieCai, Tianbao Yang and Xun Zhou: A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big data and cognitivecomputing.

v. https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96

vi. https://en.wikipedia.org/wiki/Particulates

vii. http://aqicn.org/city/indiahttps://app.cpcbccr.com/AQI_India/

viii. https://archive.ics.uci.edu/ml/data sets/Air+quality

ix. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu .” Detection and Prediction of Air Pollution using Machine Learning Models”. International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018

x. Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie.” Air Quality Prediction: Big Data and Machine Learning Approaches”. International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018

xi. https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

xii. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/