International Journal of Civil Engineering and Technology (IJCIET)
Volume 9, Issue 7, July 2018, pp. 303–312, Article ID: IJCIET_09_07_032
Available online at http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=9&IType=7
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

© **IAEME** Publication **Scopus** Indexed

ENSEMBLE LEARNING ON FORECASTING FINE GRAINED POLLUTANT LEVELS IN AIR USING RANDOM FOREST, NAIVE BAYES, DECISION TREE ALGORITHMS

Dr. Sandhya P
Associate Professor, School of Computing Science and Engineering,
VIT Chennai, Tamilnadu, India

ABSTRACT

As particulate matter in the air can cause several kinds of respiratory and cardiovascular diseases, the air quality information predicting attracts more and more attention. Knowing these information in advance is very important to protect human from health problems. With the development of computer technology, the data we can collect is increasingly becoming fine-grained. Most important of all, they need to be analyse in real-time. However, existing methods could not meet the demand of real-time analysis. In this paper, we predict air quality based on a Ipython implementation of random forest algorithm. First, a distributed random forest algorithm is implemented using Ipython on the basis of resilient distributed dataset and shared variable. Then, we build an air quality prediction model using the parallelized random forest algorithm. The proposed method is evaluated with real meteorology data obtained from IIT Madras. The experiment results show that the proposed method is fast in predicting concentration level of PM2.5. And the results also prove the effectiveness and scalability of our method when deal with big data.

Key words: Decision Tree, Ensemble learning, Fine-grained pollutant level, Naïve Bayes, Random Forest.

Cite this Article: Dr. Sandhya P, Ensemble Learning on Forecasting Fine Grained Pollutant Levels in Air using Random Forest, Naive Bayes, Decision Tree Algorithms, International Journal of Civil Engineering and Technology, 9(7), 2018, pp. 303–312.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=9&IType=7

1. INTRODUCTION

Air pollution has attracted more and more attention in recent years, especially in fast growing urban cities in developing countries where air contaminant becomes part of people's daily experience. According to WHO, air pollution is a major environmental risk to health. By reducing air pollution levels, countries can reduce the burden of disease from stroke, heart disease, lung cancer, and both chronic and acute respiratory diseases, including asthma. The lower the levels of air pollution, the better cardiovascular and respiratory health of the

population will be, both long- and short-term. Therefore, it is imperative that we have the ability to forecast pollutant levels so that local authorities can issue health alert more effectively and design area specific pollution control measures accordingly.

In every pollutant monitoring study, fine grained particulate measurements are mentioned alongside the gaseous pollutants. In general particulate matter in atmosphere is driven by natural activity and the most common source of particulate matter is still oceanic salt sprays. Other natural sources are volcanic activity, storms, forest and grassland fires etc. In recent times, human activities like increasing usage of fossil fuels has led to significant jump in anthropogenic aerosols (those made by human activity) and it now accounts for about 10% of total atmospheric aerosols.

The term fine grained pollutants refers to the particles in atmosphere having size smaller than 2.5$_{\mu}$m. In our analysis and forecasting, we are focused on PM2.5 (Particulate Matter with less than 2.5$_{\mu}$m diameter) since they are especially dangerous with a 36% increase in lung cancer per 10 $_{\mu}$g /m3 beyond current safety standards. Since they are so small and light, they stay longer in air and this increases the chances of inhaling by humans and animals. PM2.5 are able to bypass the nose and throat and penetrate deep into the lungs and some may even enter the circulatory system. Long-term exposure to PM2.5 may lead to plaque deposits in arteries, causing vascular inflammation and a hardening of the arteries which can eventually lead to heart attack and stroke.

Scientists in the study estimated that for every $10\mu g/m^3$ increase in fine particulate air pollution, there is an associated 4%, 6% and 8% increased risk of all-cause, cardiopulmonary and lung cancer mortality, respectively [1,2,3]. The goal of our project is to construct a predictive model for PM2.5 across USA using input parameters such as temperature, wind, pressure, etc. Our ideal output is the PM2.5 level of the next 24 hours given historical conditions.

2. PROPOSED SYSTEM:

In this project we will build predictive model for PM2.5 levels based on historical conditions by giving set of input data's like Temperature, Pressure, Relative Humidity, Wind Speed [4,5].

3. ALGORITHMS USED

The core idea of our approach is to use historical readings for physical parameters for predicting PM2.5 level of the next 24 hours [6,7,8]. To achieve our goal, we use Linear Regression, Random Forest, and Gradient Boosting Machine respectively.

- Linear Regression is most simple therefore the most best known model. It's representation is a linear equation. Making predictions is as simple as solving the equation for a specific set of inputs. Due to strict assumptions for a least squares based linear regression model, the model might not lead to a decent predictive function in a real world data-set unless we avoid a large number of pitfall carefully.
- Random Forest is one of the most widely used ensemble method. Random forests are a combination of tree predictors such that each tree in the ensemble is built from a sample drawn with replacement from the training set. As the number of trees grows, the generalization error for forests becomes less and less and eventually converges. In fact, in this project, Random Forest is the most predictive model which outperforms both Linear Regression as well as Gradient Boosting Machine, with an error rate cluster around 0 within 20% of the real value excluding outliers.

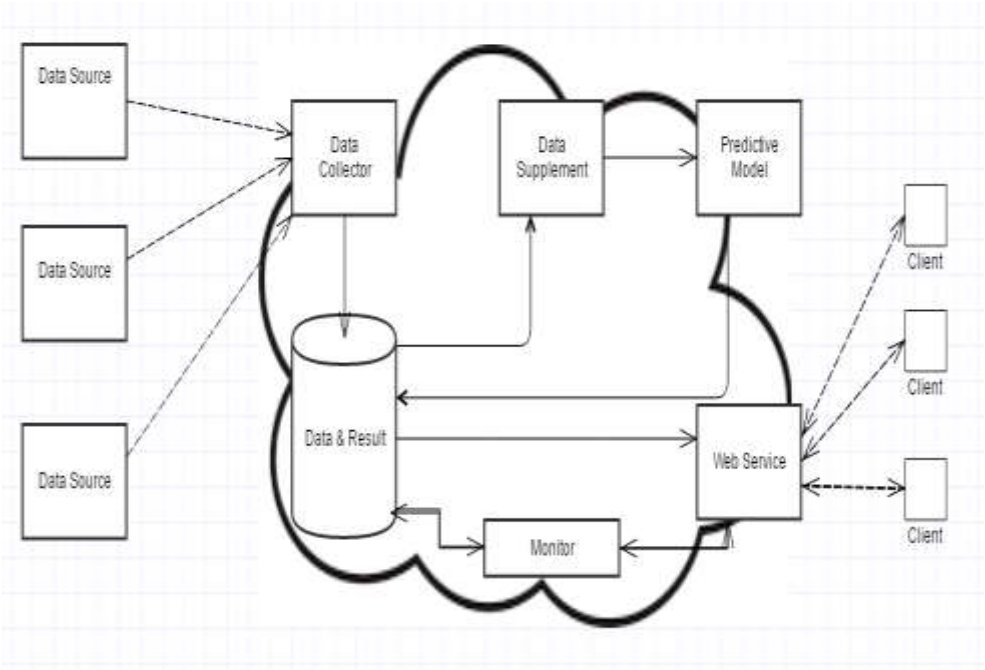

Figure 1 Proposed Architecture

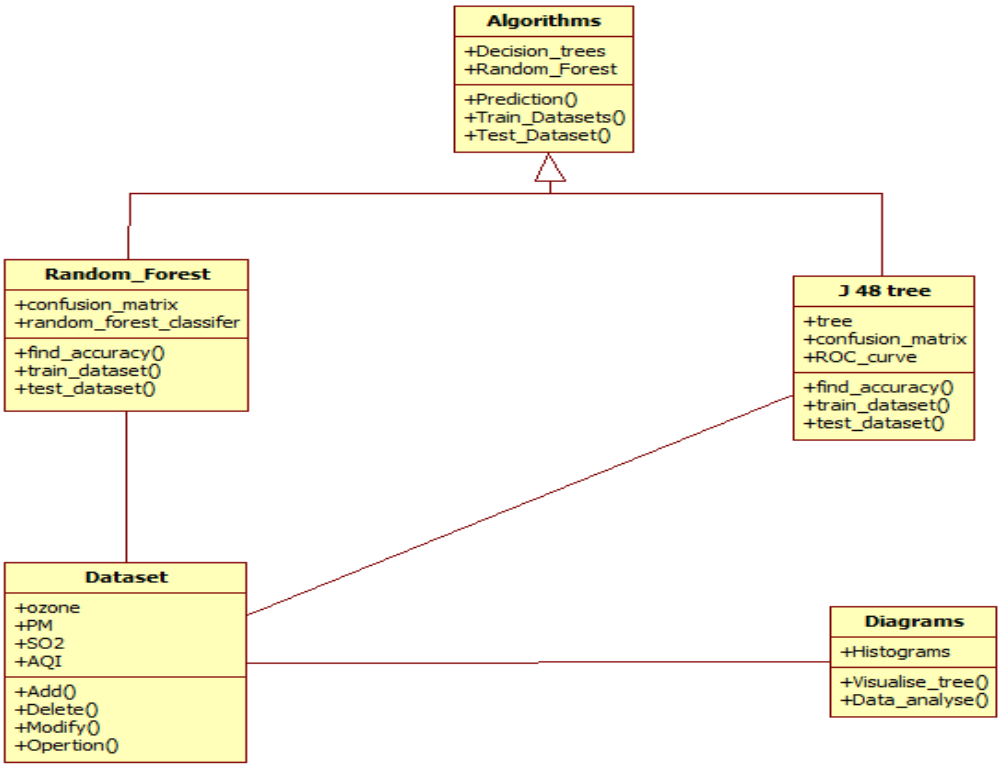

Figure 2 Class Diagram

AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5-1.0
Moderately polluted (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748	17-34	801-1600	1200-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

AQI	Associated Health Impacts
Good (0-50)	Minimal impact
Satisfactory (51-100)	May cause minor breathing discomfort to sensitive people.
Moderately polluted (101–200)	May cause breathing discomfort to people with lung disease such as asthma, and discomfort to people with heart disease, children and older adults.
Poor (201-300)	May cause breathing discomfort to people on prolonged exposure, and discomfort to people with heart disease.
Very poor (301-400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases.
Severe (401-500)	May cause respiratory impact even on healthy people, and serious health impacts on people with lung/heart disease. The health impacts may be experienced even during light physical activity.

Figure 3 AQI – Category, Pollutants and Health Breakpoints

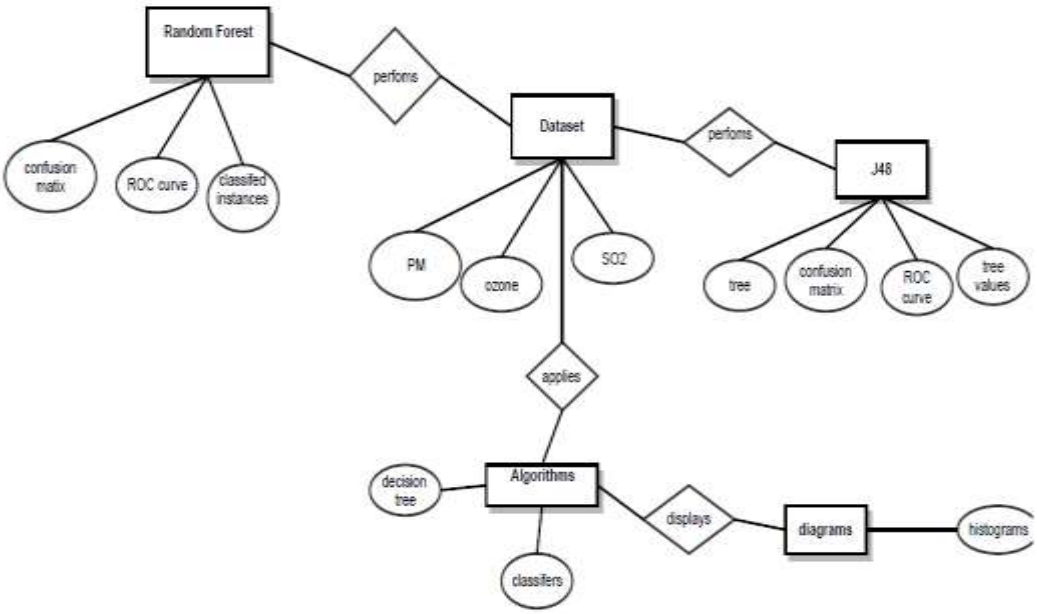

Figure 4 ER Diagram

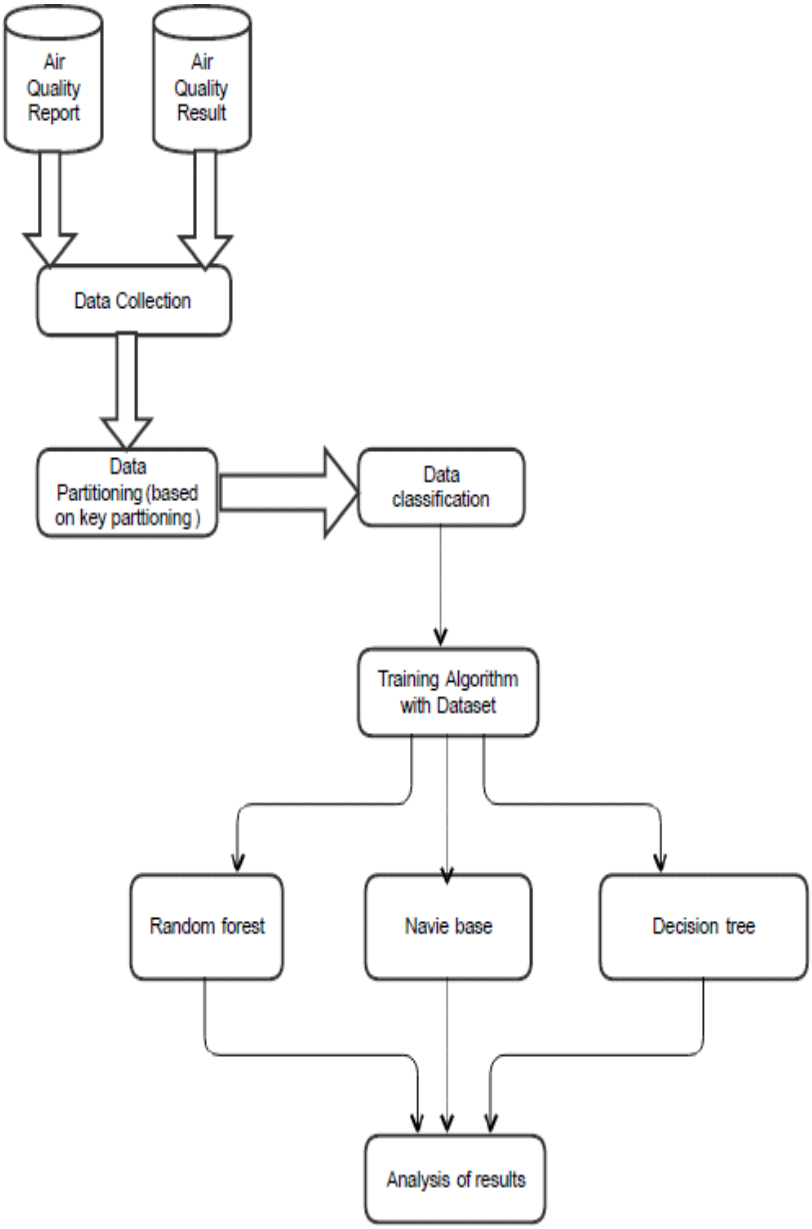

Figure 5 Data Flow Diagram

4. ALGORITHMS

The core idea of our approach is to use historical readings for physical parameters for predicting air quality level of the next 24 hours. To achieve our goal, we use Random Forest, Naïve Bayes, Decision Tree and k-means clustering Algorithms [9,10].

4.1. Random Forest

It is one of the most widely used ensemble method. Random forests are a combination of tree predictors such that each tree in the ensemble is built from a sample drawn with replacement from the training set [11,12].

Step 1: Should import all the following packages.

```
import pandas as pd

import numpy as np
```

```
import scipy as sp
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.cross_validation import cross_val_score
from sklearn.metrics import roc_auc_score
```

Step2: Load train and test data .

```
train = pd.read_csv("/home/narendra/2017/jan1.csv")
test = pd.read_csv("/home/narendra/2017/sample.csv")
```

Step3: To plot histograms

```
train.Ozone.hist()
plt.title('Histogram of Ozone')
plt.xlabel('Ozone')
plt.ylabel('Frequency')
```

Step4: Converting data into array

```
cols = [ 'Ozone','SulfurDioxide','PM']
colsRes = ['Quality']
trainArr = train.as_matrix(cols)
trainRes = np.ravel(train.as_matrix(colsRes))
```

Step5: RandomForestClassifier

```
rf = RandomForestClassifier(n_estimators=1000) # initialize
rf.fit(trainArr, trainRes)
```

Step6: Prediction of results

```
testArr = test.as_matrix(cols)
results = rf.predict(testArr)
test['predictions'] = results
test
```

Step7: Prediction Score

```
rf.score(testArr,predicted)
```

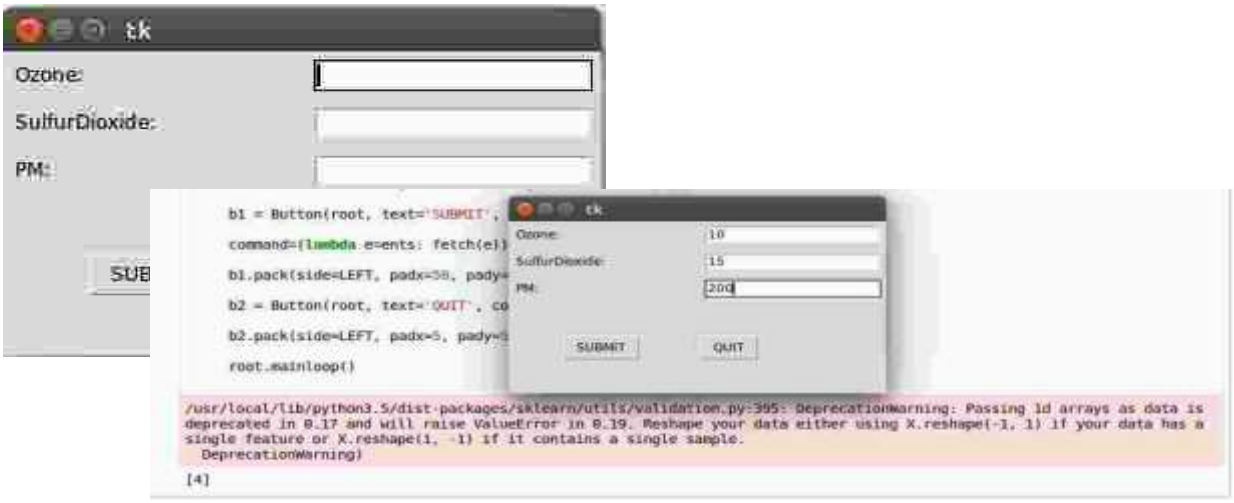

Figure 6 Random Forest

4.2. Decision Algorithm

--Decision Tree Classifier--

clf = DecisionTreeClassifier()

clf.fit(trainArr,trainRes)

testArr = test.as_matrix(cols)

results= clf.predict(testArr)

results

4.3 Naïve Bayes Algorithm:

-- Navie Base Classifier--

clf = GaussianNB()

clf.fit(trainArr,trainRes)

testArr = test.as_matrix(cols)

results= clf.predict(testArr)

results

5. CALCULATIONS OF AQI

The general equation for the sub-index (I_i) for a given pollutant concentration (C_p) [13,14,15]

Ip = [(Ihi-Ilow)/(BPhi-BPlow)] (Cp-BPlow)+Ilow

Where,

BPhi ->Breakpoint concerntration greater or equal to given concerntration

Bplow ->Breakpoint concerntration smaller or equal to given concerntration

Ihi -> AQI value corresponding to Bphi

Ilow -> AQI value corresponding to Bplo

Cp -> Pollutant concentration

Formula

=IF(ISTEXT(D2),0,IF(D2<=30,D2*50/30,IF(AND(D2>30,D2<=60),50+(D2-30)*50/30,IF(AND(D2>60,D2<=90),100+(D2-60)*100/30,IF(AND(D2>90,D2<=120),200+(D2-90)*(100/30),IF(AND(D2>120,D2<=250),300+(D2-120)*(100/130),IF(D2>250,400+(D2-250)*(100/130)))))))

=IF(ISTEXT(B2),0,IF(B2<=50,B2*50/50,IF(AND(B2>50,B2<=100),50+(B2-50)*50/50,IF(AND(B2>100,B2<=168),100+(B2-100)*100/68,IF(AND(B2>168,B2<=208),200+(B2-168)*(100/40),IF(AND(B2>208,B2<=748),300+(B2-208)*(100/539),IF(B2>748,400+(B2-400)*(100/539)))))))

=IF(ISTEXT(C2),0,IF(C2<=40,C2*50/40,IF(AND(C2>40,C2<=80),50+(C2-40)*50/40,IF(AND(C2>80,C2<=380),100+(C2-80)*100/300,IF(AND(C2>380,C2<=800),200+(C2-380)*(100/420),IF(AND(C2>800,C2<=1600),300+(C2-800)*(100/800),IF(C2>1600,400+(C2-1600)*(100/800)))))))

6. RESULTS

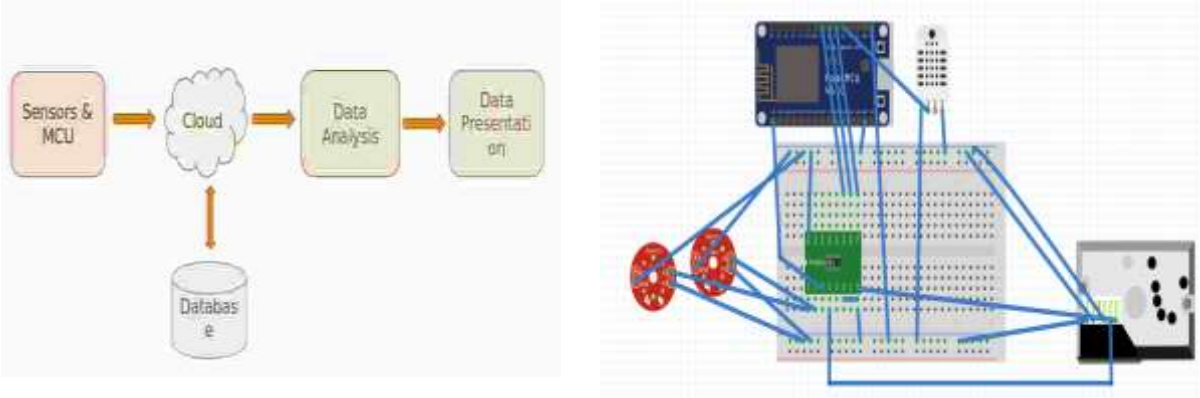

Figure 7 Block Diagram

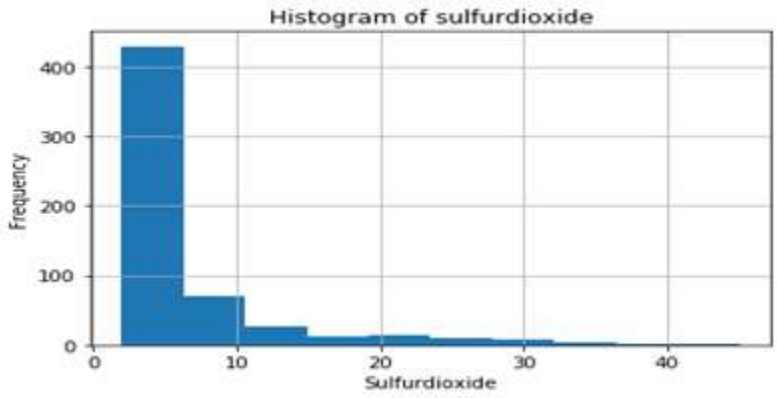

Figure 8 Histogram of sulfurdioxide

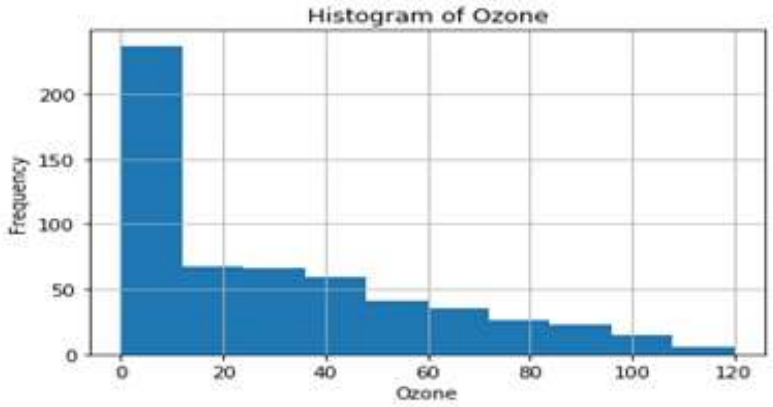

Figure 9 Histogram of Ozone

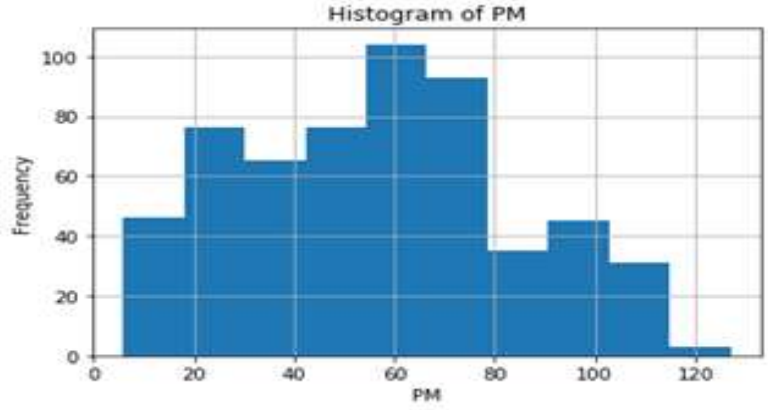

Figure 10 Histogram of PM

Table 1 Predictions

	Date	SulfurDioxide	Ozone	PM	predictions
0	08-01-2017	2.39	76.34	46	2
1	08-01-2017	2.19	79.04	46	2
2	08-01-2017	2.00	81.24	37	2
3	08-01-2017	2.32	71.07	37	2
4	08-01-2017	2.41	88.10	37	2
5	08-01-2017	2.44	77.81	37	2
6	08-01-2017	2.28	69.10	43	2
7	08-01-2017	2.18	79.79	43	2
8	08-01-2017	2.44	85.34	43	2
9	08-01-2017	2.51	86.56	43	2
10	08-01-2017	2.52	84.14	27	2
11	08-01-2017	2.54	88.51	27	2
12	08-01-2017	2.91	82.57	27	2
13	08-01-2017	3.07	52.18	27	2

7. CONCLUSIONS

In this paper, we report on a real-time air quality forecasting system that uses data-driven models to predict fine-grained air quality over the following 48 hours. The system is based on a framework that connects the cloud with clients, collecting air quality, meteorological data

and weather forecasts from over CPCB in India. We evaluate our predictive method with data from Chennai, presenting the results of two major places: IIT Madras and Alandur.

REFERENCES

[1] Urban Air Website: http://urbanair.msra.cn/

[2] Urban Air Windows Phone Client: http://www.windowsphone.com/en-us/store/app/urban-air/f36d5a33-2ccc-45f5-afd2-0c1afc5fc6dc

[3] Air Quality Forecasting on Bing Map: http://cn.bing.com/ditu/

[4] Air Quality Research Subcommittee of the Committee on Environment and Natural Resources CENR. Air Quality Forecasting: A Review of Federal Programs and Research Needs, June 2001

[5] Environmental Protection. Guideline for Developing an Ozone Forecasting Program. EPA-454/R-99-009. July 1999

[6] Hsieh, H. P., Lin, S. D., Zheng, Y. Inferring Air Quality for Station Location Recommendation Based on Big Data. In Proc. of KDD 2015, 2015.

[7] Air Pollution Forecasting in the UK, http://www.airquality.co.uk/archive/reports/ list.php.

[8] Lewis, R. J. An Introduction to Classification and Regression Tree (CART) Analysis. Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine.

[9] Shang, J., Zheng, Y., Tong, W., Chang, E. and Yu, Y. "Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City," In Proc. of KDD'14, pp. 1027-1036, 2014.

[10] Vardoulakis, S., Fisher, B. E. A., Pericleous, K., Gonzalez-Flesca, N. Modelling Air Quality in Street Canyons: A Review. Atmospheric Environment 37 (2003), pp. 155-182.

[11] Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., Real-time Air Quality Forecasting, Part I: History, techniques, and current status, Atmospheric Environment 60 (2012), pp. 632–655.

[12] D. Steinberg and P. Colla, "CART: classification and regressiontrees," The top ten algorithms in data mining, vol. 9, p. 179, 2009.

[13] J. Han, M. Kamber, and J. Pei, Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.

[14] Z.-H. Zhou, Ensemble methods: foundations and algorithms. CRC Press, 2012.

[15] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, and M. G. Institute, "Big data: The next frontier for innovation, competition, and productivity,"2011.

[16] Mahadev, Vinod Kumar and Himani Sharma, Detection and Analysis of DDOS Attack at Application Layer Using Naïve Bayes Classifier. International Journal of Computer Engineering & Technology, 9(3), 2018, pp. 208–217.