Air Pollution Prediction System for Smart city using Data Mining Technique

Heni Patel[1], Swarndeep Saket[2]

[1]*Student of Masters of Engineering, Ahmedabad, Dept. of Computer Engineering, L. J. Institute of Engineering & Technology, Gujarat, India*

[2]*Assistant Professor, Ahmedabad, P G Dept. of Computer Engineering, L. J. Institute of Engineering & Technology, Gujarat, India*

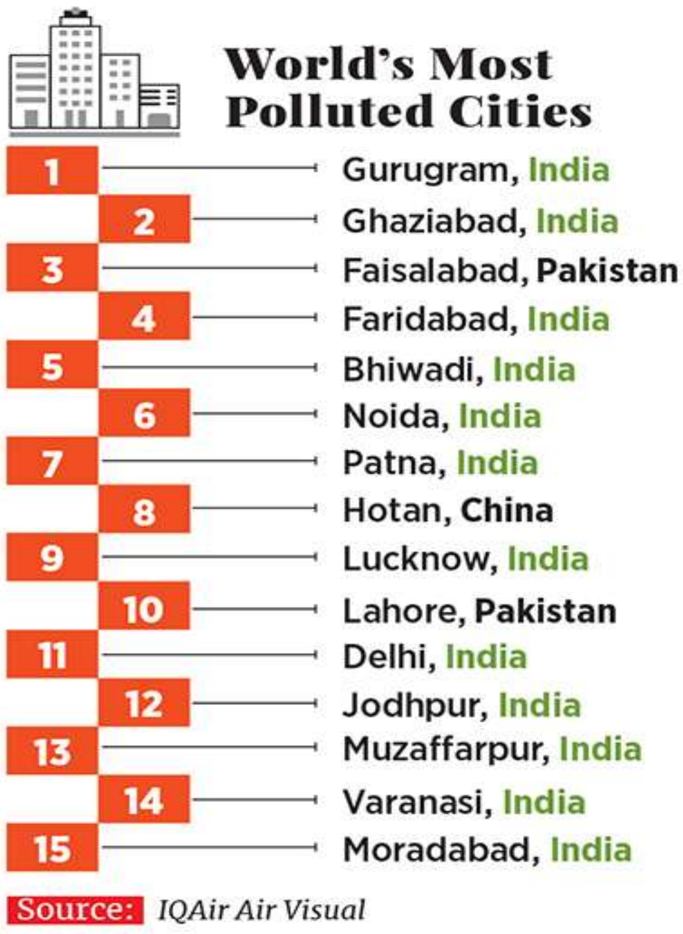
---***---

Abstract - *Each living organism needs fresh and good quality air for every second. None of the living things can survive without such air. But today air pollution becomes one of the major hazards. Because of automobiles, agricultural activities, factories and industries, mining activities, burning of fossil fuels our air is getting polluted. These activities spread sulphur dioxide(SO2), nitrogen dioxide(NO2), carbon monoxide(CO), Ozone(O3), particulate matter(PM10 &PM2.5) pollutants in our air which is harmful for all living organism. The air we breathe every moment causes several health issues. So we need a good system that predicts such pollutions and is helpful to make our environment better. It leads us to look for advance techniques for predicting such pollutants. So here we are predicting air pollution for our smart city using data mining technique. In our model we are using multivariate multistep Time Series data mining technique using random forest algorithm. Our system takes time series data of these pollutants. Also takes data of temperature, wind speed & direction and applies them to our model to predict air pollution. This model reduces the complexity and improves the effectiveness and practicability and can provide more reliable and accurate decision for environmental protection departments for smart city.*

Key Words: Air pollution prediction, Data mining, Smart city, Time series, Random Forest Algorithm, Complexity, Effectiveness, Practicable.

1. INTRODUCTION

In 2017, 1.2 million people died in India due to pollution, which is 12.5% of total deaths within our country. Also almost 2 lakh children lost their lives because of air pollution-related diseases, which suggests on an average 535 deaths occurred daily. The Union Health Ministry and Indian Council of Medical Research says in our country one out of each eight deaths is attributed to pollution.



Source: IQAir Air Visual

Fig-1: World's Most Polluted cities[10]

As Figure1.1 shows World's most polluted cities are in India. So pollution is one most important hazards among all environmental pollution. As each living organism must needs air for each second. So Air should be fresh and of good quality. None of the living things can survive without such air.

These are some reason for our air is getting polluted.

1. Combustion of fossil fuels, like coal and oil for electricity and road transport, producing air pollutants like nitrogen and gas.

2. Emissions from industries and factories, releasing great amount of monoxide, hydrocarbon, chemicals and organic compounds into the air.

3. Agricultural activities, thanks to the utilization of pesticides, insecticides, and fertilizers that emit harmful chemicals.

4. Waste production, mostly because of methane generation in landfills.

These activities spread pollutants in our air which is harmful for all of us. Information about those pollutants are given below.

1. Particulate Matter (PM2.5 and PM10): Living in a city, you've likely walked outside your apartment and noticed a layer of grey haze that prevents you from clearly seeing the landscape miles ahead. That haze appears when there are high concentrations of particulate matter (PM) in the air.PM is a mixture of solids or liquid droplets in the air that are categorized by size:

PM10**:** Inhalable particles that are less than or equal to 10 micrometers in diameter. Examples include dust, pollen, and mold.

PM2.5**:** Fine particles that are less than or equal to 2.5 micrometers in diameter. To put this in perspective, they are about 1/30th of a strand of human hair (too small for the human eye to see).

2. Nitrogen Dioxide (NO2): Another dangerous form of urban pollution may be a group of gases called nitrogen oxides. because of burning of fuel from road vehicles, cookers and heaters gas is produced during warmth. They react within the air to make particulate (PM) and ozone.

3. Sulphur Dioxide (SO2): Sulfur dioxide may be a gas. It invisible and incorporates a nasty, sharp smell. About 99% of the dioxide in air comes from human sources. The most source of sulphur dioxide is industrial activity that processes materials that contain sulfur, example. The generation of electricity from gas, oil or coal that contains sulfur. SO2 affects human health. It irritates the nose, throat, and airways to cause coughing, wheezing, shortness of breath, or a good feeling round the chest.

4. Carbon Monoxide (CO): Carbon monoxide is colorless and odorless, but highly toxic. While often thought of as an inside hazard, it is also a significant outdoor pollution further. Main sources of CO to outdoor air are cars, trucks and other vehicles or machinery that burn fossil fuels.

5. Ozone (O3): Ground-level ozone (also known as the "bad" ozone) is created by a chemical reaction in the presence of sunlight that forms between man-made VOCs and nitrogen oxides This explains why ozone levels tend to be higher and subsequently more dangerous within the summertime. In rural areas, downwind of urban areas or industrial sites, the highest levels of ozone are mostly found.

Today many health issues are happening due to the air. Our air become more polluted day by day, which is extremely bad for all the living organism. Not only these pollutants but also there are some more parameters which affect the air pollution. Some parameters which affect the air pollution which is explained below.

1. Temperature
Temperature inversion have a great effect on air quality. Increase of height in atmosphere makes air cooler. However sometimes an upper air layer is warmer than lower air layer. That's call inversion in atmosphere. And if inversion stays for long time on the ground than pollutants can build up to the higher level. That is how temperature affect the air pollution.

2. Wind direction and speed
Measurement of wind speed and wind direction is vital in air quality monitoring. It can help identify the situation of the source of the pollution, and also provide a higher overall picture of what's happening within the air.

3. Air quality of previous day
Previous days quality of air must affect the following day.so here we are using the pollution level of the previous day. So we are able to get the accurate result.

So these are the essential of the air pollutants and a few factors that are affect the pollution. Using of these attribute here we are predicting the pollution. Also now we see the motivation and objective about our project.

2. DATA MINING FOR PREDICTION

Data mining is a process which is used to turn raw data into useful information. Data mining is the process of gathering information and analyzing it for actionable patterns, which may then be employed in many things. There are many uses of data mining technique. One usage of data mining technique is for prediction. At the end we extract useful information from the data.

Here is basic flow diagram of knowledge discovery from data.

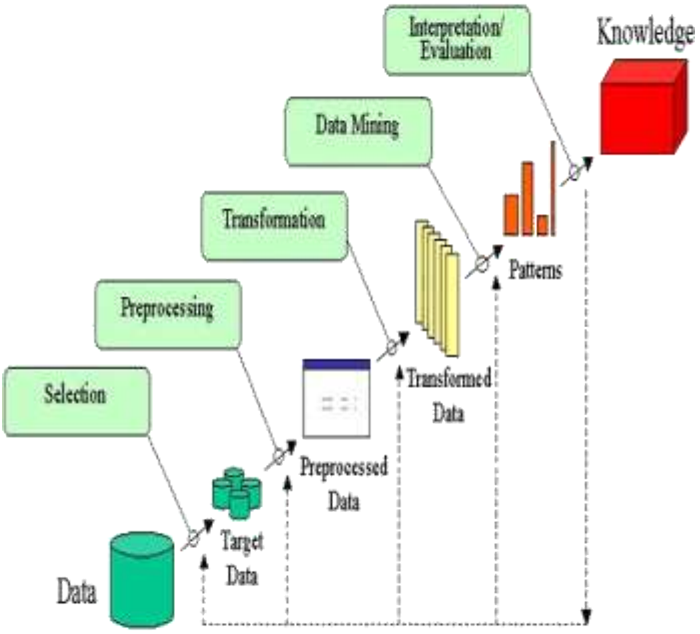


Fig-2: knowledge discovery process using data mining[18]

Steps of KDD process:

- Data Cleaning - In this step the noise and inconsistent data is removed.
- Data Integration - In this step multiple data sources are combined.
- Data Selection - In this step relevant to the analysis task are retrieved from the database.
- Data Transformation - In this step data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining - In this step intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation - In this step, data patterns are evaluated.
- Knowledge Presentation - In this step, knowledge is represented.
- Here are the main types of Data mining algorithms.

Time series forecasting: Time series algorithms are the same as regression algorithms in this they predict numerical values but statistic is concentrated on forecasting future values of a criterion.

Time series data points are image of the past. Understanding historical events, patterns and trends are some basic indicators that each one businesses track. Nowadays, with the ever-growing amount of "big data" and therefore the need for near real-time insights, statistics analysis is becoming a necessary a part of business decisions.

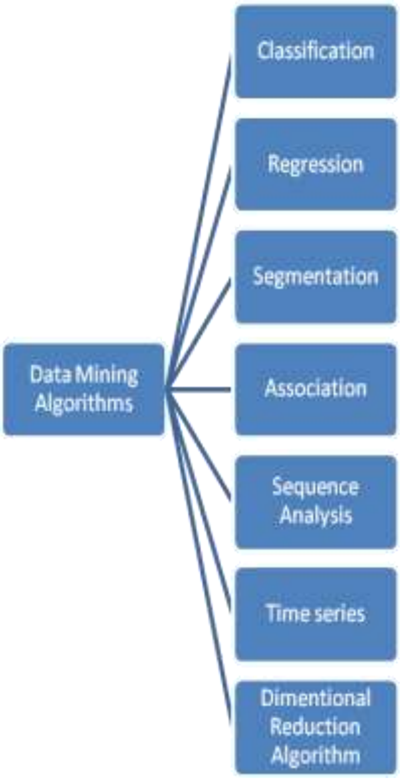


Fig-3: Basic Data Mining Algorithms[3]

Time series forecasting could be a significant a part of data mining and machine learning technologies that involve fitting machine learning models to form predictions. These tools are as simple because the extrapolation of historical trends into the longer term.

2.1. MULTIVARIATE MULTISTEP TIME SERIES PREDICTION USING RANDOM FOREST

Multivariate time series has more than one **time-**dependent variable. Each variable depends not only on its past values but also has some dependency on other variables. This dependency is used for forecasting future values predicting multiple time steps into the future is called multi-step time series forecasting. Making predictions about the future is called extrapolation in the classical statistical handling of time series data. More modern fields focus on the topic and refer to it as time series forecasting[17].

Components of Time Series:

Analysis of the time series provides a body of techniques for better understanding a dataset. Perhaps the most useful of these is the decomposition into 4 constituent parts of a time series:

- **Level -** The baseline value for the series if it were a straight line.
- **Seasonality** - Repeating patterns or cycles of behavior over time.

- **Noise** - Variability in the observations that cannot be explained by the model.
- **Trend** - Linear increasing or decreasing behavior of the series over time.

Random Forest Algorithm is a supervised algorithm for classification. As the name suggests, this algorithm creates a number of trees into the forest. In general, the more trees that appear in the forest, the more robust the forest is. Similarly, the higher the number of trees in the forest in the random forest classifier yields the high precision tests. So here we'll use this whole technique to make prediction of air pollution in smart cities[17].

3. RELATEDWORK

In this section, we discuss the different papers related to air pollution prediction technique. We take all the recent years papers.

Shweta Taneja, Dr. Nidhi Sharma, Kettun Oberoi, Yash Navoria, proposed paper of Predicting Trends in Air Pollution in Delhi using Data Mining. In this Paper, They have used time series analysis method for analyzing the pollution trends in Delhi and predicting about the future. The time series method includes Multilayer Perceptron and Linear Regression[1].

In Elseviere(2018) paper, Forecasting air pollution load in Delhi using data analysis tools. In this paper,A systematic approach has been followed in this analysis. The approach starts with the collection of dataset from CPCB. Collected data has been pre processed to remove the redundancy. Pre processing of data includes steps like parsing of dates, noise removal, cleaning, training and scaling. Further, descriptive analysis has been carried out on two different platforms-Rstudio and Tableau for different stations. For observing the forecasted results, predictive analysis has been done[2].

KRZYSZTOF SIWEK, STANISŁAW OSOWSKI, Proposed paper for Data Mining methods for Prediction of Air Pollution.The paper will discuss the numerical aspects of the air pollution prediction problem, concentrating on the methods of data mining used for building the most accurate model of prediction.In this paper feature selection is done by using the genetic algorithm (GA). The application of several predictors and feature selection methods allowed integrating their results into one final forecast. The best results of integration were obtained in the direct application of selected features to the RF, performing at the same time the role of regression and integration[3].

In Springer (2019) Paper, Prediction of Air Quality Using Time Series Data Mining. Many of the modern databases are temporal, which makes the task of studying and developing time series data mining techniques an important and much needed task. Time series data mining identifies time-dependent features from time series databases. These features are used for building predictive models. This paper proposes an efficient algorithm to predict the concentration of the various air pollutants by using time series datamining techniques. The time series datamining algorithm CTSPD or Continuous Target Sequence Pattern Discovery has been used for the prediction of air pollutants. The predictions made by the proposed solution are compared with the predictions made by SAFAR-India and found that the proposed solution provides more accurate results. By studying the obtained air quality patterns, it was found that the concentration of a pollutant need not depend on all the other pollutants[4].

In Springer (2018) Paper, Air Pollution Prediction Using Extreme Learning Machine: A Case Study on Delhi. In this work multi-variable linear regression model of ELM is used to predict air quality index for PM10, PM2.5, NO2, CO, O3. In the proposed model, the previous day air quality index of pollutants and meteorological conditions are used for prediction. Performance of the proposed model was compared with the prediction of an existing prediction system, SAFAR as well as with the actual values of next day. ELM-based prediction was found to have greater accuracy than the existing[5].

Khaled Bashir Shaban, Senior Member, IEEE, Abdullah Kadri, Member, IEEE, and Eman Rezk Proposed Paper of Urban Air Pollution Monitoring System With Forecasting Models. In this paper air quality data are collected wirelessly from monitoring motes that are equipped with an array of gaseous and meteorological sensors. These data are analyzed and used in forecasting concentration values of pollutants using intelligent machine to machine platform. The platform uses ML-based algorithms to build the forecasting models by learning from the collected data. These models predict 1, 8, 12,nd 24 hours ahead of concentration values. Based on extensive experiments, M5P outperforms other algorithms for all gases in all horizons in terms of NRMSE and PTA because of the tree structure efficiency and powerful generalization ability. On the other hand, ANN achieved the worst results because of its poor generalization ability when working on small dataset with many attributes that leads to a complex network that overfit the data, while having SVM better than ANN in our case due to its adaptability with high dimensional data[6].

Table -1: Comparison Table

PUBLICATION	TITLE	METHOD	LIMITATION
IEEE, 2016	Predicting Trends in Air pollution in Delhi using Data	Linear regression, Multilayer perceptron, Time series	Linear regression only looks at linear relationships

	Mining.	analysis	between dependent and independent variables. Sometimes this is incorrect.
IEEE, 2016	Air Pollution Monitoring System With Forecasting Models.	SVM(Support Vector Machine),ANN(Artificial neural Network)	Neural Networks requires filling missing values and converting categorical data into numerical. we need to define the NN architecture. How many layers to use, usually 2 or 3 layers should be enough. How many neurons to use in each layer? What activation functions to use? What weights initialization to use?
AMCS,2016	Data mining methods for prediction of air pollution	SVM Regression RF_fusion	SVM algorithm is not suitable for large data sets. SVM does not perform very well, when the data set has more noise.
Springer, 2018	Pollution prediction using extreme learning machine: a case study on delhi.	ELM(Extreme Machine Learning)	ELM is much faster to train, but cannot encode more than 1 layer of abstraction, so it cannot be "deep".
Elseviere, 2018	Forecasting air pollution load in Delhi using data analysis tools.	Time series regression	Here we are using time series with regression.so again regression is limited to the linear relationship. it is easily affected by outliers.
Springer, 2019	Prediction of Air Quality Using Time Series Data Mining.	CTSPD(Continuous Target Sequential Pattern Discovery)	The main drawbacks Sequential pattern mining in particular, are: the large amount of discovered patterns; its inability to use background knowledge; and the lack of focus on user expectations.

4. PROPOSED WORK

Fig-4 shows the block diagram of the proposed model. Which shows the basic flow of our system and steps we do.

Step1 Collection of Data: There are many sensors available at many places which sense the pollutants. Using one of those here we are collecting the air pollutant data.

Step 2 Preprocessing of Data: Preprocessing is a step where we remove noise and filling the missing data using some appropriate values.

Step 3 Feature Selection: Feature selection is the process of finding the most relevant inputs for predictive model. This technique can be used to identify and remove unneeded, irrelevant and redundant features that do not contribute or decrease the accuracy of the predictive model.

Step 4 Multivariate Multistep Time Series Prediction Using Random Forest: At this stage we take multivariate time series data and we predict air pollution using random forest algorithm. There are multiple trees and each tree is trained on a

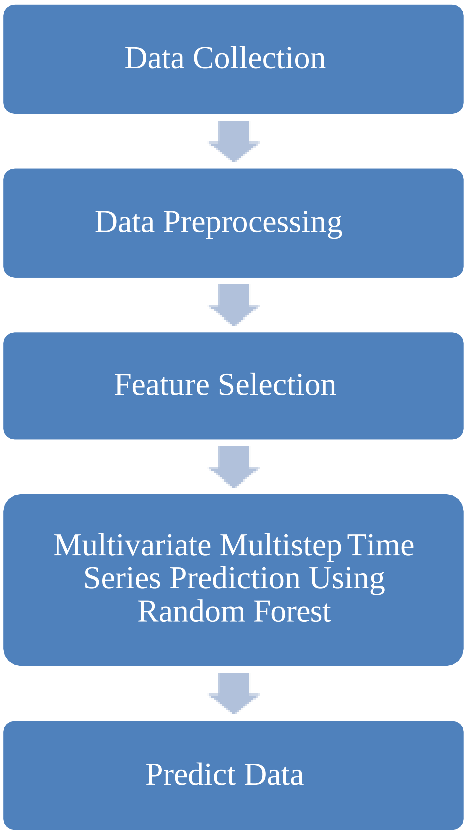

Data Collection
Data Preprocessing
Feature Selection
Multivariate Multistep Time Series Prediction Using Random Forest
Predict Data

Fig-4: Workflow of proposed method for Air Pollution Prediction[4]

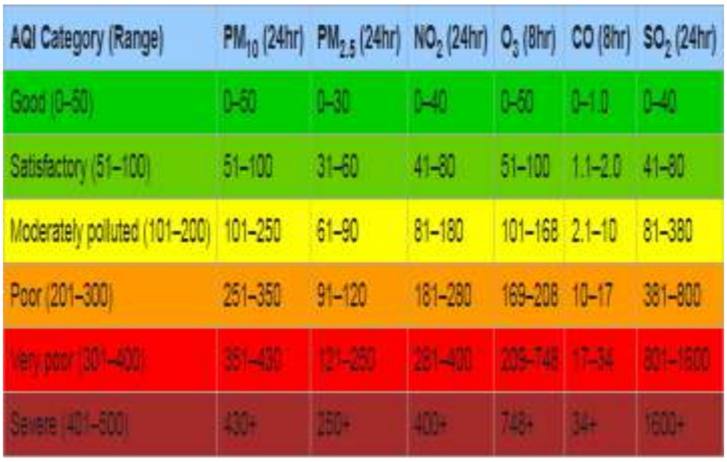
subset of time series data. It predicts the value using all of those decision trees. And we're doing recursive predictive strategy. Using this recursive predicting strategy here we are getting multiple results at a time.

In which this model shown widely. First original data is converted into small samples. In this we are doing this using bootstrap sampling processing. This process makes same size of samples every time. Using all this samples model makes decision trees from this. All decision trees make decision or predict the value. And here we repeatedly doing this process to make multiple prediction.

Step 5 Prediction: here our system gives result of predicted value. We use the predicted value for the next prediction. So this is how our model predict the pollutants.

5. IMPLEMANTATION

Here we are predicting all the pollutants for the next day. We are predicting PM2.5, PM10, SO2, NO2, Ozone, CO pollutants. We are predicting data using the base model and the proposed model. Also we are showing the mean square error and mean absolute error of both model. Below are the screenshots of prediction of all the pollutants.



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

Fig-5: AQI Measurement Table

We are showing our result in a color bar, in which color of bar is showing quality index of pollutant.

5.1 Prediction of PM2.5

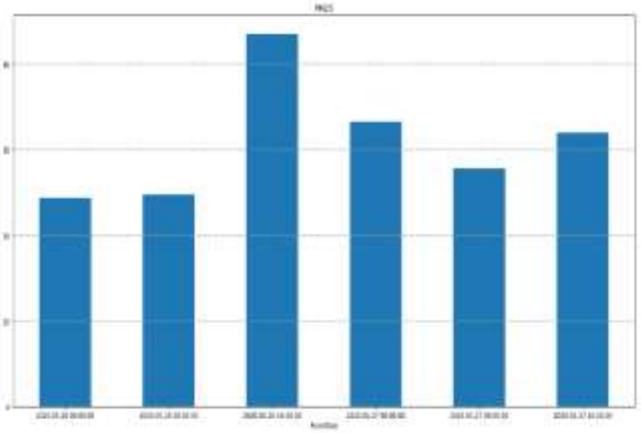


Fig-5.1.1: PM2.5 data

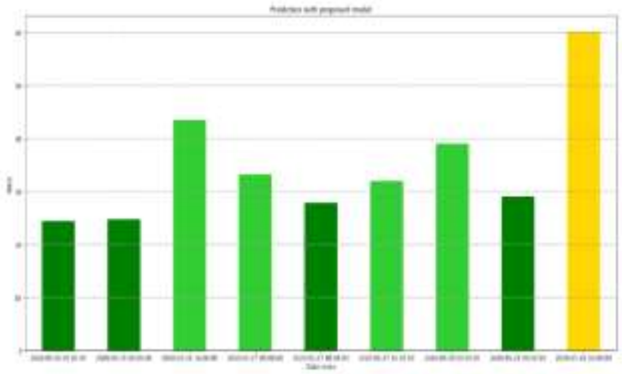


Fig-5.1.2: Prediction of PM2.5

5.2 Prediction of PM10

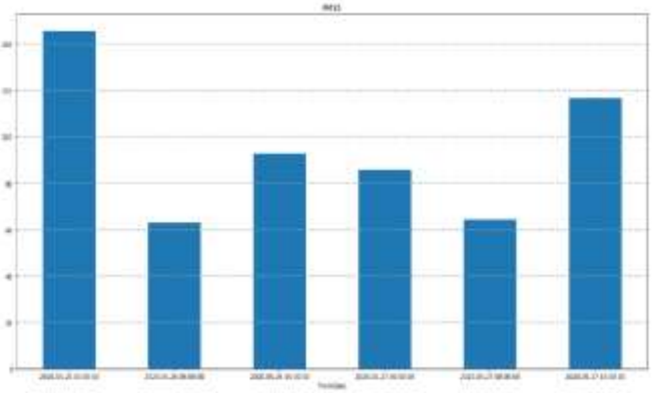


Fig-5.2.1: PM10 data

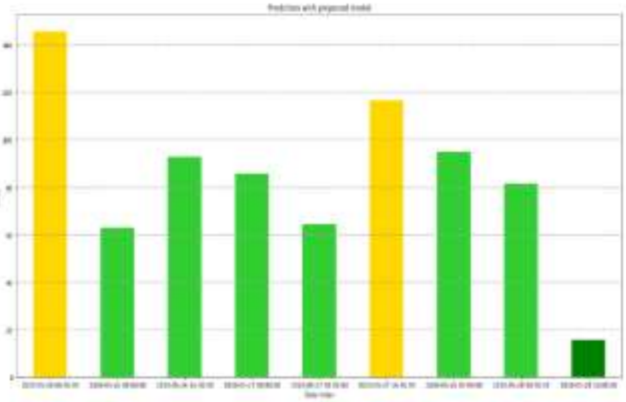


Fig-5.2.2: Prediction of PM10

5.3 Prediction of SO2

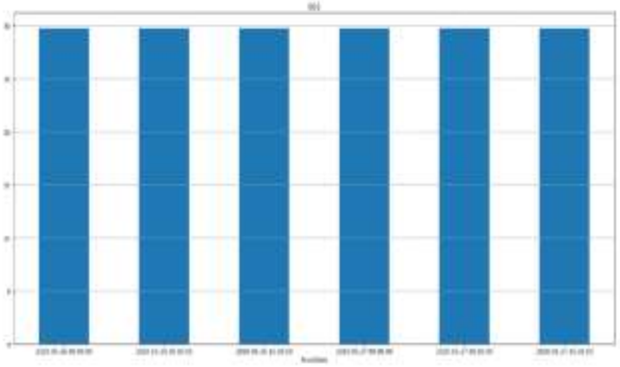


Fig-5.3.1: SO2 data

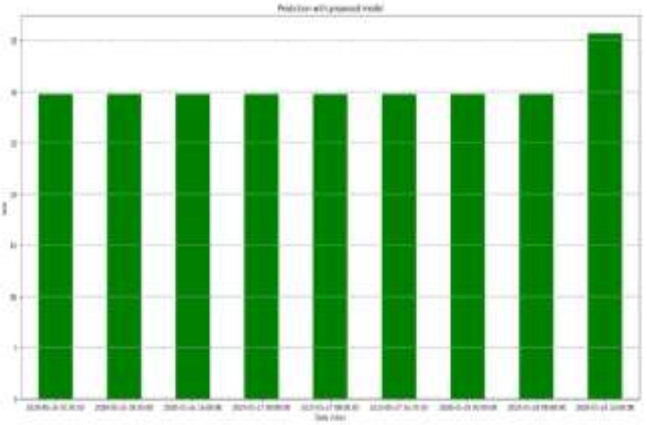


Fig-5.3.2: Prediction of SO2

5.4 Prediction of NO2

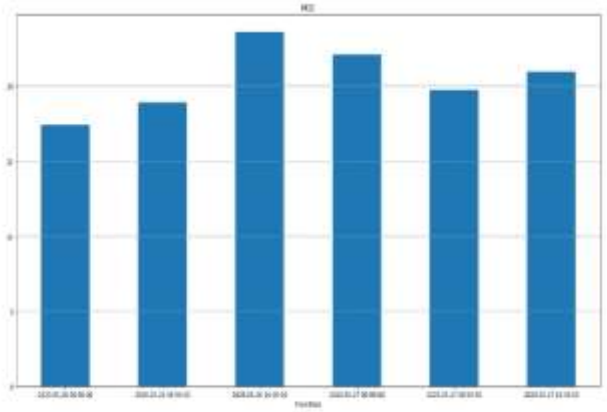


Fig-5.4.1: NO2 data

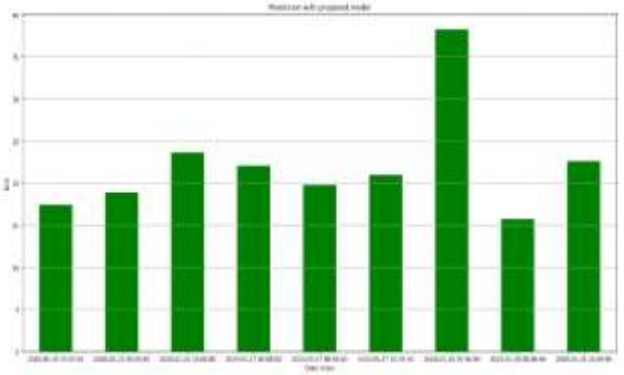


Fig-5.4.2: Prediction of NO2

5.5 Prediction of Ozone

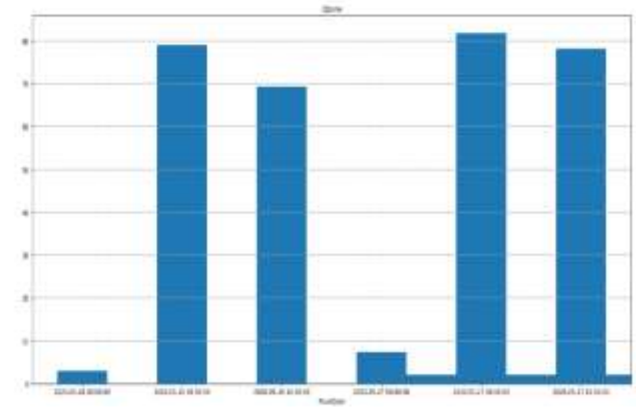


Fig-5.5.1: Ozone data

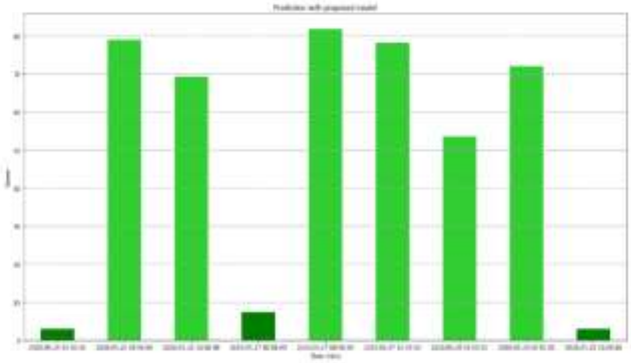


Fig-5.1.2: Prediction of Ozone

5.6 Prediction of CO

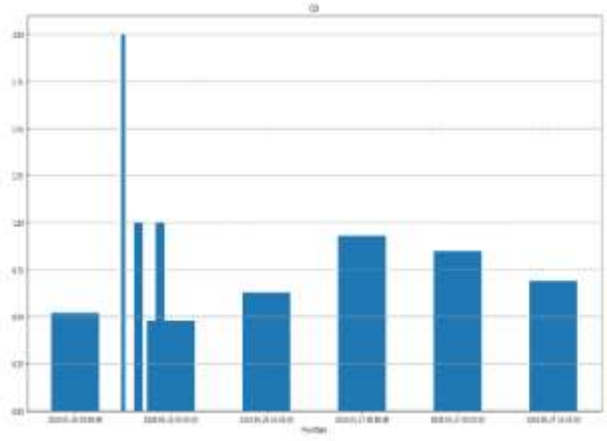


Fig-5.5.1: CO data

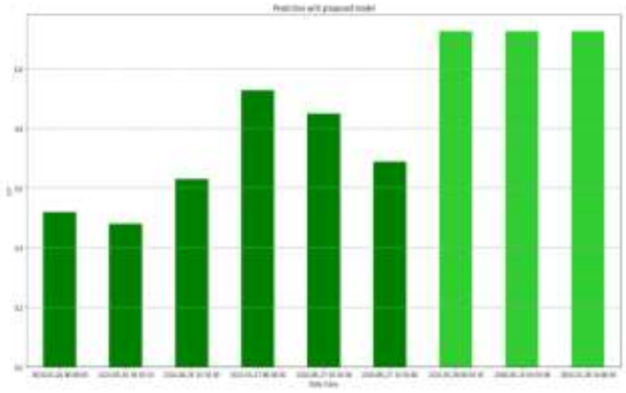


Fig-5.5.2: Prediction of CO

6. CONCLUSIONS

Our system helps in improving the prediction of air pollution in our smart city. Multivariate Multistep Time Series Prediction Using Random Forest technique improve the performance and reduce the complexity of the air pollution prediction model. Also here we are using feature selection technique, which make our prediction even better.

Here we are predicting PM2.5, PM10, SO2, NO2, Ozone, Co pollutants. Also we shown prediction in a bar chart. We give color to the bars, which indicates the quality index of the pollutants. That the pollutant level is good, moderate or bad.

Using this System we are trying to spread awareness of the air pollution in these smart cities, so that everyone can live healthy and happy life. Also together we will make our India clean and pollution free.

REFERENCES

[1] Shweta Taneja, Dr. Nidhi Sharma, Kettun Oberoi, Yash Navoria ,"Predicting Trends in Air Pollution in Delhi using Data Mining", IEEE(2016)

[2] Nidhi Sharmaa, Shweta Tanejab*, Vaishali Sagarc, Arshita Bhattd, "Forecasting air pollution load in Delhi using data analysis tools.", Elseviere (ICCIDS 2018)

[3] KRZYSZTOF SIWEK, STANISŁAW OSOWSKI," Data mining methods for prediction of Air Pollution", amcs(2016)

[4] Mansi Yadav, Suruchi Jain and K. R. Seeja," Prediction of Air Quality Using Time Series Data Mining", Springer (2019)

[5] Manisha Bisht and K.R. Seeja," Air Pollution Prediction Using Extreme Learning Machine: A Case Study on Delhi.", Springer(2018)

[6] Khaled Bashir Shaban, Senior Member, IEEE, Abdullah Kadri, Member, IEEE, and Eman Rezk," Urban Air Pollution Monitoring System With Forecasting Models.", IEEE(2016)

[7] Khaled Bashir Shaban, Abdullah Kadri, Eman Rezk ,"Urban Air Pollution Monitoring System With Forecasting Models",IEEE SENSORS JOURNAL, VOL. 16, NO. 8, APRIL 15, 201

[8] Forecasting Criteria Air Pollutants Using Data Driven Approaches; An Indian Case Study Tikhe Shruti, Dr. Mrs. Khare , Dr. Londhe ,IOSR-JESTFT (Mar. - Apr. 2013)

[9] Forecasting Criteria Air Pollutants Using Data Driven Approaches; An Indian Case Study Tikhe Shruti, Dr. Mrs. Khare , Dr. Londhe,IOSR-JESTFT (Mar. - Apr. 2013)

[10] Air Quality Forecasting Methods,

"https://www.airvisual.com/air-pollution-information/research/air-quality-forecast-methods"

[11] Multivariate Multistep Time series Forecasting model for Air pollution. "https://machinelearningmastery.com/how-to-develop-machine-learning-models-for- multivariate-multi-step-air-pollution-time-series-forecasting/K. Elissa"

[12] Yi-Ting Tsai, Yu-Ren,Zeng, Yue-Shan Chang, "Air pollution forecasting using RNN with LSTM", IEEE(2018)

[13] Min Huang, Tao Zhang, Jingyang Wang and Likun Zhu," A New Air Quality Forecasting Model Using Data Mining and Artificial Neural Network", IEEE,(2015)

[14] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, hengqiang 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

[15] Ebrahim Sahafizadeh, Esmail Ahmadi," Prediction of Air Pollution of Boushehr City Using Data Mining", 2009 Second International Conference on Environmental and Computer Science.

[16]About Jupyter notebook,"https://www.infoworld.com/article/3347406/what-is-jupyter-notebook-data-analysis-made-easier.html"

[17] Survey paper published in dissertation project 1," https://www.irjet.net/archives/V6/i12/IRJET-V6I12159.pdf"

[18] KDD flow diagram," http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html"

[19] KDD process steps," https://data-flair.training/blogs/data-mining-and-knowledge-discovery/"