Prediction of Spatial-Temporal PM2.5 using Non-Parametric Technique

Madhuri.N[1]

[1]MSRIT Institute of Technology, MSRIT post, Bangalore [Bangalore, 560054,India]

---***---

Abstract: *Air pollution can directly affect human beings, people are surrounded by highly polluted air that contains PM2.5. Very fine particles which are less than 2.5 microns penetrate deeper into the respiratory system and cause health issues. Due to high industrial activities, there is an increase in PM2.5. To meet necessary accurate air quality, we propose a non-parametric technique such as random forest, linear regression, SVR, and random ferns. Statistical analysis of data, missing value treatments, validating data, and data cleaning is performed on the air dataset. We consider hourly data on (PM2.5) provided by the central pollution control board (CPCB), We use machine learning approaches. The Random Fern method is being demonstrated to be effective in predicting PM2.5. Later ensemble all the models using boosting technique for better accuracy.*

Keywords: *Regression, Air quality, Non-parametric Technique, Ensemble, Boosting, Accuracy.*

1.INTRODUCION

Air quality index (AQI) depends on the estimation of particulate matter, (PM) particulate matter is a mixture of many small dust particles and water droplets these are emitted from power plants, forest fires, automobiles, and due to industrial activities, there is an increase in particulate matter. (PM) includes particles such as pollen grains from seeds and flowers and smoke from vehicles. PM2.5 particles are generally 2.5 micrometres' which are smaller in diameter they penetrate deeper into the lungs and cause health issues. It's important to predict PM values and precise prediction provides valuable information to protect mankind from being damaged by pollution; so, we consider daily hourly data on (PM2.5) from the central pollution control board (CPCB), meteorological parameters are being considered such as temperature, barometric pressure, wind speed, wind direction, humidity, we use a random forest technique during training phase many decision trees constructions occur. Majority of trees voting is chosen as the final decision. We also use a technique such as bootstrap aggregation, Linear regression, SVR, and another machine learning approach called Random Ferns.

2.Related work

Medhin Zamsni and Chuniaxng Cao [1] have experimented on "PM2.5 prediction based on random forest XG boost and deep learning using multisource remote sensing data", here authors have studied importance of PM2.5 prediction in Tehran's urban area, and they have applied few machines learning approaches like extreme XG boost, random forest, they have used all 23 features such as satellite, ground measured PM2.5 etc. they have calculated the best model performance and finally they used the DNN model for prediction and XG boost has performed better in predicting PM value. Huang, Q xiao [2] carried performance analysis on "Predicting monthly high- resolution PM2.5 concentration with Random forest model in the north China plain" authors have developed ml model on monthly level for estimating PM2.5 levels in north China plain they have used Random forest approach of multi-angle implementation of atmospheric correction(MAIAC) and aerosol optical depth (AOD) they have taken 2013 and 2015 land cover and ground PM2.5 measurement they found that southern Hebei, and western northern Henan were most polluted areas, and the model can predict historical PM2.5 concentration at seasonal, annual, and monthly based levels.

W. Yuchi, E. Gombojav [3] author carried performance analysis to predict indoor Particulate matter concentration for the pregnant women's they were part of portable air cleaners in magnolia "Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in the highly polluted city" they have used models such as random forest model and(MLR) multiple linear regression they have also developed a model which combines all models were evaluated in 10-fold cross-validation and finally the blended MLR model which includes RFR prediction had the better performance. Jing Wei, Wei Huang "Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach" [4] authors have experimented that space-time random forest can predict monthly, daily and annual concentration of PM2.5 and accuracy across China, this is out performance population models STRF is robust it is an accurate estimator of PM2.5 concentration and by taking added advantage of regression ensemble model they got a high resolution and high- quality results. Xhu, Jhbelle, Xmeng [5] authors have "Estimated PM2.5 concentration in the conterminous United States using random forest approach", they have incorporated aerosol optical depth (AOD)and few meteorological fields and land use data they have also used random forest ensemble learning methods to predict high accuracy finally the result predicts cross-validation r2 score of 0.80 and the root mean squared prediction error of 1.78.

Ryu, Y.yang and G Han [6] authors have done experiment analysis on “a random forest approach for predicting air quality in urban sensing systems”, they have generated urban sensing data and they have used road information and real-time traffic status and meteorology data is being evaluated with actual data of the city ,They implemented the random forest algorithms for predicting uncovered regions AQI prediction is of 81% the result has outperformed performance single decision tree, naïve Bayes, and ANN. Chiou-Jye Huang, ping-Huan Kuo [7] authors experimented preparing “ A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities” they have estimated PM2. 5concentration, using methods such as convolutional neural network (CNN), and other method called long short-term memory (LSTM) are combined and applied to PM2.5 forecasting system. Finally, CNN-LSTM model and its feasibility and for forecasting pm value are being verified in this paper and this technology is useful in improving the ability to estimate the air pollution in the smart cities.

Y Zhan, Y Luo, and X Deng [8] authors have experimented on “spatial-temporal prediction of daily ambient ozone levels across China using the random forest for human exposure assessment”, they have implemented approach called random forest to predict daily maximum concentration across China in 2015 and finally they found that this study was the first statistical modelling work of ambient O3 for China at the national level and the dataset which they have used is valuable for rarefy the epidemiological analysis on O3 pollution in China . Chang, Xhu Y Liu [9]authors have experimented “calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentration via statistical scaling study” explains public available satellite retrieved AOD data is additional covariate for predicting daily PM2.5 prediction the main objective of the paper is to develop a statistical model that utilizes they have used down scaling methodology for predicting PM2.5 and ozone concentration. Here the accuracy is improved by spatial-temporal regression coefficient. Finally they have overcome the challenge of spatial misalignment between the point of monitoring measurement and the gridded AOD data they have predicted the PM2.5 concentration at any point within the grid cell.TV VU, Z Shi, j Cheng [10] authors have experimented on “assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique”, they have implemented the random forest modelling and Theil-Sen regression technique results indicate that in reducing primary pollution finally they have implemented the PM10, PM2.5, SO2, and CO levels have been decreased the average concentration of O3 has been increased by 4.9%.

3.Proposed System Architecture

In System Architecture, we collect the air pollution data from (CPCB) central pollution control board, then the data is being pre-processed and data is split into 80% training data and 20% testing data. Later we apply the machine learning algorithms and train the model finally we validate the model and predict the result.

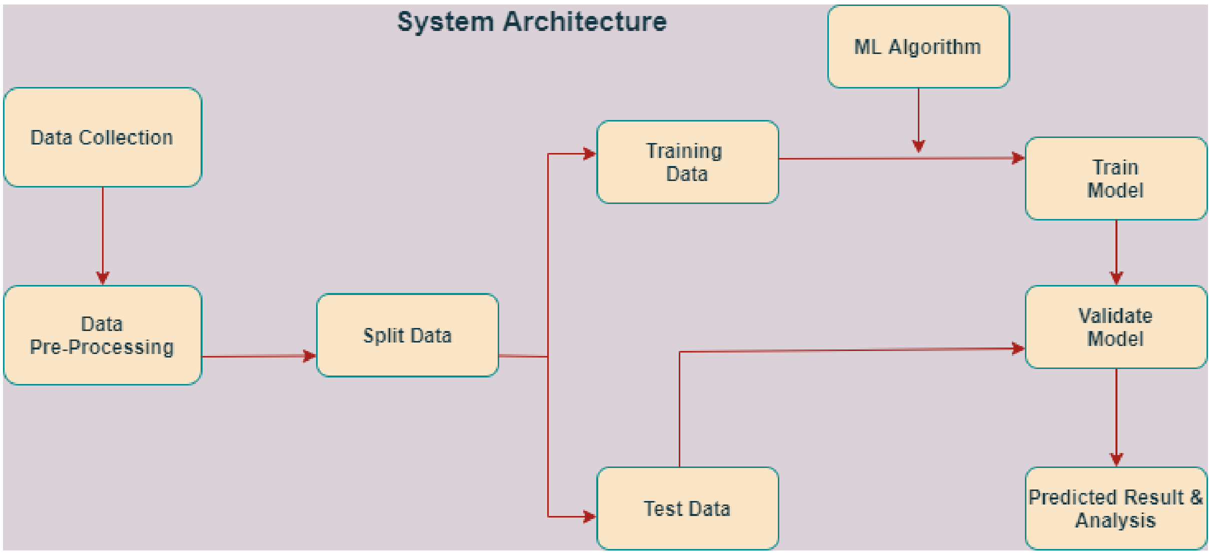


Fig 1: System Architecture.

- **Data collection**

The First step is to collect all the necessary data such as relative humidity, barometric pressure, temperature, the velocity of wind speed, wind direction these are the meteorological parameters we collect all this information from the central pollution board. (CPCB) is an organization under the ministry of environment whose main aim is to prevent and control pollution. we also consider road information of Bangalore city.

- **Exploratory Data Analysis**

EDA steps help in understanding the structure and correlation of our data. Basically, it is an approach to analyse a dataset, often we can use data visualization methods and graphs.

- **Data Pre-processing**

This involves data cleaning and few transformation methods are being used to remove the outliers so that we can easily create a model, we have used a method called MICE computation which is "Multiple imputations by chained equations"

- **Model training**

After constructing different machine learning models its essential to train the models, after constructing models, model can recognize data, later split the data set into 80% training data set and 20% testing dataset, later we can build train data using training dataset.

- **Construction of predictive Models**

We have created 4 models called a random forest, random fern, linear regression, and SVR we split data into train and test later we perform X is fixed number of 80% of training data and the rest 20% is test data these models are predicting effectively with minimal error later ensemble all the model using boosting technique to get better accuracy.

3.1 Data Description

We have initially collected the data from (CPCB)central pollution control board in general, the more data we have the better is the result predicted. We collect Data such as relative humidity, barometric pressure, temperature, the velocity of wind speed, wind direction; these are the meteorological parameters. We have considered hourly data of BTM and Jayanagar stations five years dataset from 2015 to 2020.

We have considered the parameters such as Temp, WD, PM2.5, RH, WS, BP, VWS

(a). Relative humidity: The amount of water vapor present in the air is defined as Relative humidity, it has a greater effect on increasing PM2.5 the relative humidity ranges between 65-80 percent. If relative humidity increases then PM2.5 decreases.

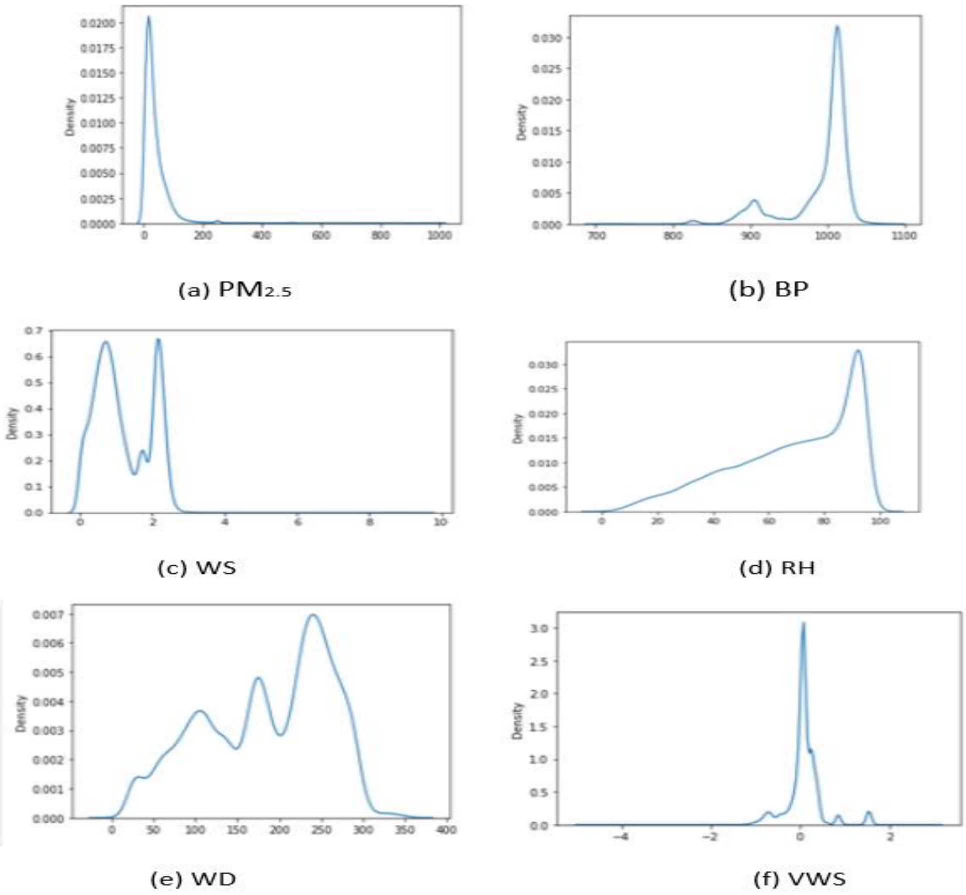
(b). Temperature: The Temperature has a strong positive correlation with PM2.5 , If temperature increases PM2.5 will increase, vice versa if temperature decreases the PM2.5 decreases. It ranges between 32 degrees Celsius.

(c). Velocity of wind speed (VWS): An increase in wind speed will increase PM2.5 concentration, velocity of wind speed is measured through an anemometer. It ranges from 1.15mph (meter per hour),0.51 m /sec.

(d). Wind direction: PM2.5 has a weak correlation with wind direction which is -0.1, wind direction has a strong positive correlation with pressure it ranges from 0.501 m/s.

(e). Wind speed: If wind speed is lower than 3 m/s we have a +ve correlation with PM2.5, if the speed of wind is greater than 3 m/s it can transport numerous pollutants from faraway it ranges from 1.15 mph (mile per hour).

(f). Particulate matter 2.5 (PM2.5): PM2.5 is a small air-borne particle that will penetrate deeper into the lungs and cause health issues. Unhealthy PM2.5 ranges from 151 to 200 which is 55.5 to 150.4.



(a) PM2.5

(b) BP

(c) WS

(d) RH

(e) WD

(f) VWS

Fig 2: plots of all the parameters.

These kernel density plots help us to visualize the data distribution overtime period, the kernel density plot which is also known as KD plot these are the smoother versions of the histogram. Mainly they are good representations of data or the numerical values.

3.2 Data Pre-processing

Sometimes data is unstructured, the reason behind the missing data might be data is not being collected continuously, mistakes in entering the data, and few technical problems with biometers. For cleaning the data, we have used the MICE technique which is "Multiple imputations by chained equations ". "Multiple Imputation" this is a strong informative methodology of managing missing information in the dataset. The procedure "fills in" (imputes) missing information in every iteration, every such variable within the data set through an iterative series of prophetical models. In every iteration, every such variable within the dataset is imputed exploitation of the opposite variables within the dataset. These are the steps involved in cleaning the Data Cleaning, Data Integration, Data Transformation, Data Reduction.

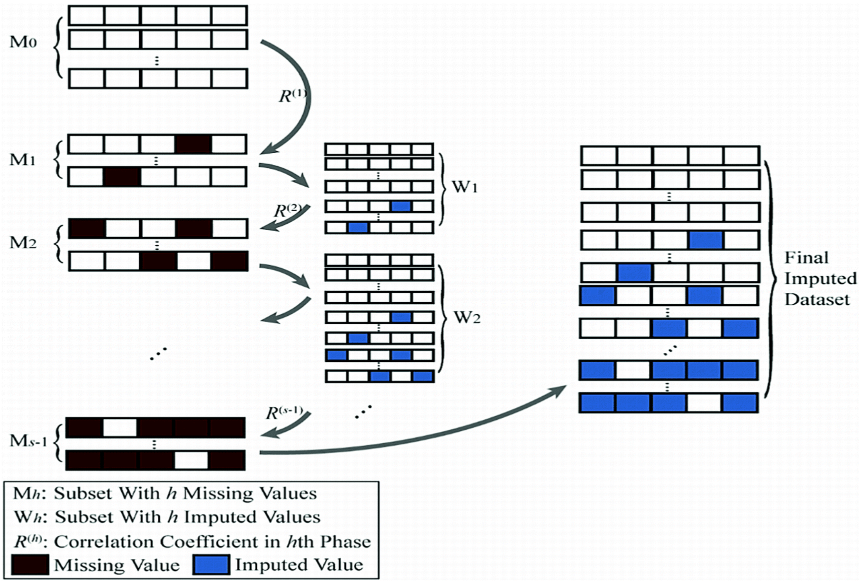

Mo
M1
M2
Ms-1
R(1)
R(2)
R(s-1)
W1
W2
Final Imputed Dataset
Mh: Subset With h Missing Values
Wh: Subset With h Imputed Values
R(h): Correlation Coefficient in hth Phase
■ Missing Value 🟦 Imputed Value

Fig 3: MICE (Multiple imputations by chained equations)

3.3 PM2.5 prediction using Random Forest and Random Ferns.

In random forest technique, during training phase many decision trees constructions occur. Majority of trees voting is chosen as the final decision. Huge number of trees operate as one committee and finally they end up giving final decision. Random ferns are a simple and powerful method, The random fern algorithm is decision tree ensemble. They have extended features of the random forest, The objective behind this approach is to create a simple and efficient algorithm and there is a similar connection to decision tree and ensembles like a random forest.

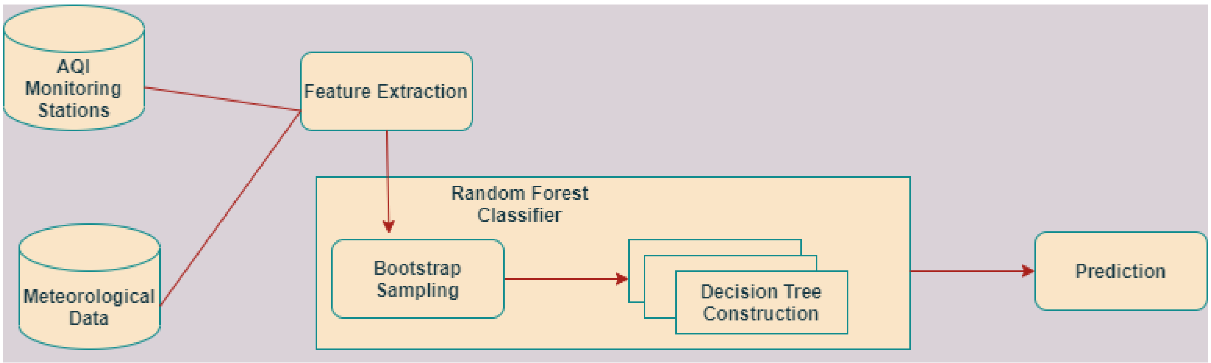
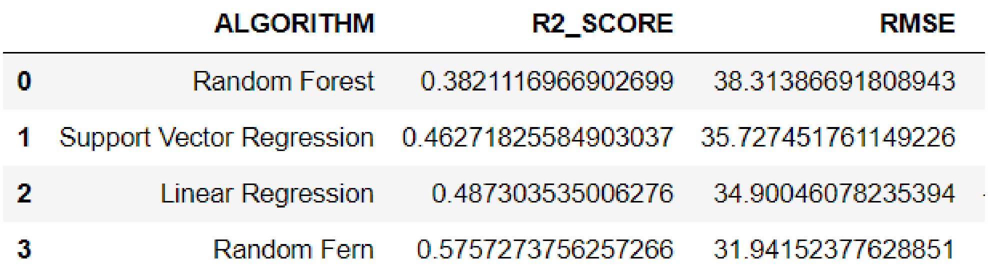


Fig 4:PM2.5 prediction using random forest

Fern is a constrained tree where the same binary test is performed at each level of a tree a fern is defined as a simplified binary decision tree, random ferns are easy to understand and this will provide a probabilistic output and it will appear to outperform random forest. And training time grows linearly with fern size.

4.Experiments and Results

We are using four different models such as Random Forest, Random fern, Linear Regression, and SVR. Random ferns are a simple and powerful method, it is considered as a decision tree ensemble. Random fern method is in this random forest technique, during training phase many decision trees constructions occur. Majority of trees voting is chosen as the final decision. extended with corresponding features of the random forest algorithm. linear regression is a linear approach which will model the relationship between one or more explanatory variables. We have also applied SVR which is a support vector regressor. It is a supervised machine learning model that helps in analysing data for regression; it is built on the concept of support vector regression. These above are the four models which are used to predict spatial and temporal PM2.5. These models are predicting effectively with minimal error, Random fern model with the aid of comparing the great accuracy, gives better accuracy prediction among all models later we ensemble all the model using boosting technique to get better accuracy.



	ALGORITHM	R2_SCORE	RMSE
0	Random Forest	0.3821116966902699	38.31386691808943
1	Support Vector Regression	0.46271825584903037	35.727451761149226
2	Linear Regression	0.487303535006276	34.90046078235394
3	Random Fern	0.5757273756257266	31.94152377628851

Table 1: Accuracy of all used Models

The result of all the algorithms is measured with respect to performance, the result is calculated using RMSE value, r2 _score, the root mean square is a prediction error, RMSE is a measure that tells how the data is spread around the best fit line, r2_score is the Best possible score which is of 1.0 and this can be negative. The constant model has to predict expected value, in the above table we can see that random fern is giving less RMSE value which means compare to all other model it giving better result. Lesser is the RMSE value better is the model performance. Random ferns are giving better result boosting is a method of combining all the simple algorithms which will give an improved predictive performance. This produces a predictive model which is in the form of the week prediction model it is used to combine the weighted prediction to obtain a single weighted prediction

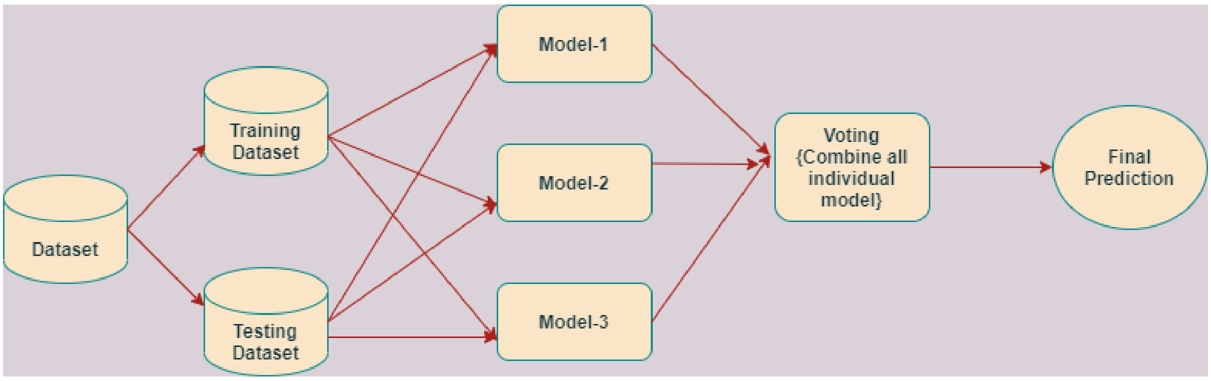
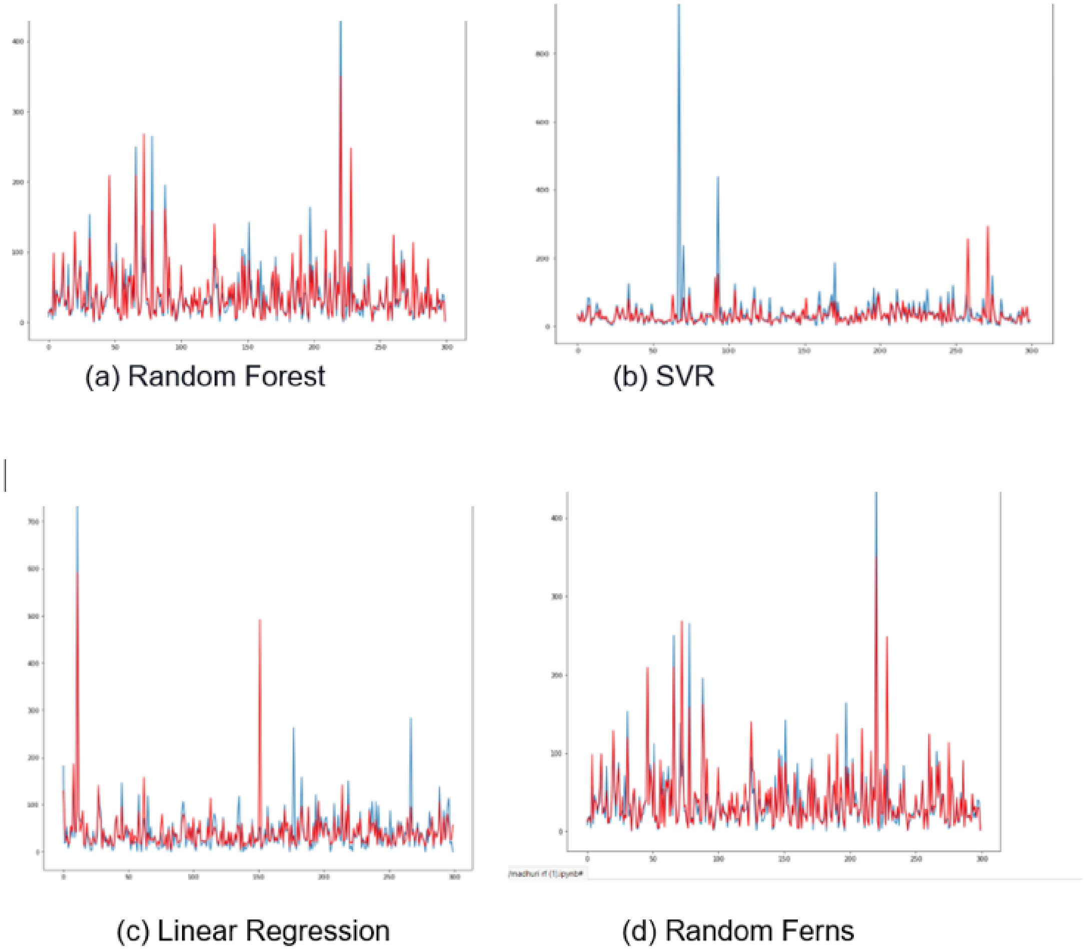


Fig 5: Voting combining individual model

Similar to numerous ML models which centre around great forecast done by a model, boosting calculations try to improve the prediction power via preparing an arrangement of fair models, each repaying the shortcomings of its types, by using voting regressor boosting is a method of combining all the simple algorithms which will give an improved predictive performance. To improve accuracy, we have combined all the algorithms. If we combine random forest, linear regression, and SVR we can see that the training score is 68 Percentage. If we combine all the four algorithms which are the random forest, random fern, linear regression, and SVR we get training score of 82 percentage by this we can consider by combining all the models using voting, we can get high performance of models training score and by combining random fern model we get better accuracy and high score by this we can conclude that ferns are an effective method of predicting the PM2.5.



(a) Random Forest

(b) SVR

(c) Linear Regression

(d) Random Ferns

Fig 6: Actual v/s predicted graphs.

The above graphs represented in Fig 7 explains the actual and predicted values of PM2.5 of all the algorithms. The blue colour represents the actual PM2.5 values and the red colour represents the predicted values in the above four graphs we can see the actual and predicted PM2.5 readings of algorithms such as Random Forest, Linear Regression, SVR, and Random Ferns, compare to all the graphs the random ferns actual and predicted values are near to each other in other algorithms the values are more varying, so we consider the random ferns are predicting better PM2.5 results.

5.Conclusion

We predicted the spatial-temporal PM2.5 using non-parametric techniques such as random forest, random fern, linear regression, and SVR; we found that compared to all machine learning algorithms random fern is predicting better accuracy results. Later we combine all the algorithms using boosting ensemble to get better accuracy, and we apply different boosting techniques like Ada boost, Gradient boost, and XG boost, we can say that if we combine all the models, we can get better accuracy, and also random fern is predicting better PM2.5 results compared to all other algorithms.

6.Acknowldgements

The authors would like to thank contributions of all other participants who were involved in completion of this work. The authors also thank reviewers.

7.References

[1] Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B. and Talebiesfandarani, S., 2019. PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. Atmosphere, 10(7), p.373.

[2] Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D. and Liu, Y., 2018. Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. Environmental Pollution, 242, pp.675-683.

[3] Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren., 2019. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environmental Pollution, 245, pp.746-753.

[4] Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L. and Cribb, M., 2019. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. Remote Sensing of Environment, 231, p.111221.

[5] Hu, X., Belle, J., Meng, X., Wildani, A., Waller, L., Strickland, M. and Liu, Y., 2017. Estimating PM2.5Concentrations in the Conterminous United States Using the Random Forest Approach. Environmental Science & Technology, 51(12), pp.6936-6944.

[6] Yu, R., Yang, Y., Yang, L., Han, G. and Move, O., 2016. RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors, 16(1), p.86.

[7] Huang, C. and Kuo, P., 2018. A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. Sensors, 18(7), p.2220.

[8] Zhan, Y., Luo, Y., Deng, X., Grieneisen, M., Zhang, M. and Di, B., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environmental Pollution, 233, pp.464-473.

[9] Chang, H., Hu, X. and Liu, Y., 2013. Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling. Journal of Exposure Science & Environmental Epidemiology, 24(4), pp.398-404.

[10] Vu, T., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S. and Harrison, R., 2019. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. Atmospheric Chemistry and Physics, 19(17), pp.11303-11314.