PM2.5 Estimation using Supervised Learning Models

Anusha Anchan, Manasa G.R.

Abstract: ***Present era of Urbanization, mechanization, and globalization has attracted more and more Air pollution problems. However, PM 2.5 (Particulate Matter) majorly present at air, having diameter below 2.5 µm. With its high concentration leading to health issues such as lung cancer, cardiovascular disease, respiratory disease etc. With respect to this, presented work approach is building of supervised learning models, XGBoost(Extreme Gradient Boosting) along with MLR(Multiple Linear Regression),RF(Random Forest) and MLP (Multilayer Perceptron) to estimate PM2.5 congregation. The air quality data of city Changping in Beijing is taken into consideration for Analaysis. The accuracy of prediction of the four approaches is measured by using contrasting discovered value verses predicted value of PM2.5 with the help of three measuring matrices. The consequences reveals that the Random Forest algorithm outperforms other data mining strategies for the considered data. Prediction of PM2.5 concentrations will assist governing bodies in warning people who are at peak risk, and taking right measures to reduce its quantity in air also to reduce its impact on human life.***

Keywords **:** ***PM2.5, XGBoost, PM10, Python.***

I. INTRODUCTION

According to WHO records, around 80 percentage of people living in towns are unshielded to air standard which contrast with guideline limit specified by standard organizations. WHO presents a data which shows that 9 out of 10 people breathe air with elevated levels of toxins. It is estimated that 4.6 million persons get killed annually because of air pollution. At present, the ratio of people living in urban cities is more compared to rural areas which will result in elevation in air pollution level. The impacts of rapid development of the total populace are resulted in the abuse and shortage of common assets, denuding of forest and particularly natural contamination.

An ongoing recording which utilizes a worldwide barometrical science model assessed that around four million annual mortality is connected to open air contamination, which is expected to twofold by 2050, generally because of anthropological insubstantial particulate matter (streamlined breadth < 2.5 µm).

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author

Anusha Anchan*, Assistant Professor in the Department of Computer Science & Engineering.

Manasa G.R., Assistant Professor in the Department of Computer Science & Engineering.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

In congruence with recent studies happening, hold a mirror to a fact that tiny particles present in the air has been exceptionally perilous on cardiopulmonary well being. An ongoing air quality investigation in major cities of the world are surpassing the WHO's prescribed degrees of 10 µg/m3.

Despite the fact that the general degrees of fine particulate contamination have been diminishing because of dynamic endeavors executed by neighborhood and national governing bodies in most recent years, certain areas in certain cities calibre of air has kept on disintegrating. Later mirrors worldwide patterns of modernized cities and mechanization. The general susceptibility towards this issue has asked the authorities to pass laws so as to forestall the air-

contamination. Leaders responsibility is to give the residents the correct data to make them mindful of air quality rates. The appropriation of air-contamination includes a mind boggling process contingent upon various variables. Truth be told, air contamination expectation, which has a non-direct dynamism, is an exceptionally troublesome errand and requires a nearby comprehension of the scattering of air toxins in the air, which includes a colossal expense. Therefore air contamination examination is considered as an imperative issue in urban administration. To accomplish the objective, proper devices should be utilized to anticipate air contamination.

This work, presents a different machine learning models to anticipate various degrees of PM 2.5 to be specific, MLR(Multiple Linear Regression), XGBoost(Extreme Gradient Boosting), RF (Random Forest) and MLP(Multilayer Perceptron.) Every one of these models are prepared with pertinent features to anticipate PM2.5. Different model exhibitions are looked at dependent on relapse execution metric MAE (Mean Absolute Error), RMSE(Root Mean Squared Error) along with R2 Score with objective of obtaining the best model for PM2.5 estimation.

II. LITERATURE SURVEY

There has been lot of work put up already related to air quality analysis. [1] Bingyue Pan et al in has performed hourly prediction of PM2.5 concentration with various models with few parameters. [2] Multivariate Analysis and spatial is carried out on particular region by David Núñez-Alonso et al. [3] Jihan Li et al came up with time series model (AR) to predict PM 2.5 concentration. They introduced a method called Kalman filter named it as AR-Kalman hybrid which could outperform earlier model and proved to be better. [4] Chunxiang Cao et al conducted a research based on meteorological features, and remote sensing AOD(Aerosol Optical Depth) data and its effect on PM2.5 prediction in a specific city in Iran. [5]

Time series analysis was performed by Simone Andréa Pozza et al of PM10 as well as PM2.5 in a specific city using model called Holt-Winter and SARIMA. [6] Rasa Zalakeviciute et al considered two sites which are placed on Quito and variance in air contamination study was carried out. Here unlike prediction, classification was performed saying if PM2.5 is high or low according to the cut off value decided by them. [7] Dixian Zhu and team formed regularized MTL problem and various optimization techniques were adopted. [8] Jiaming Shen built ARMA(Autoregressive Moving Model),SW(Stock-Watson) model and SV (Stochastic Volatility)model for the purpose of coming six hour prediction. [9] Mahesh Babu et al Built a classification study on prediction of model,checking if model has predicted right or wrong hence finding the accuracy of the models. [10] Chandana R Deshmukh et al used autoregressive model and logistic model mainly used for classification and time series prediction.

III. PROBLEM STATEMENT

Present work is aiming at building four machine learning models aimed at estimating the PM2.5 value. Also comparative study on four models performance in estimation using relevant metrics. The models used are as follows, Multiple Linear Regressor, Random Forest Regressor, XGBoost Regressor and Multilayer Perceptron. All these models are expected to train with relevant features.

IV. DATASET

Chronicled air calibre information has been taken for the city Changping present at Beijing. This data been collected spread between March 1st 2013 to February 28th 2017. The baroscopic data is coordinated with nearest climate station present at China's Meteorological Organization. Missing information are meant as NA. This dataset has 35064 records in total which comprises of air quality measures of Chanping city in Beijing. The general statistics of dataset is given in Table 1.

Table 1: Changping Air-Monitoring Data Set

Set Qualities:	Multivariate, Time-Series	**Number of records:**	35064
Feature Qualities:	Integer, Real	**Number of features:**	17
Related Work:	Regression	**Missing Value**	Yes

The set of variables present in this dataset are: **(1)**Record number. **(2)** year. **(3)** month.**(4)**day. **(5)** hour. **(6)**PM2.5 (ug/m^3) microgram per meter cube. **(7)** PM10 (ug/m^3). **(8)** SO2 (Sulfur dioxide)(ug/m^3). **(9)**NO2(Nitrogen dioxide) (ug/m^3). **(10)** CO (Carbon Monoxide) (ug/m^3). **(11)**O3 Ozone/trioxygen (ug/m^3). **(12)** Temperature (degree Celsius). **(13)** Pressure (hPa). **(14)** Dew point temperature (degree Celsius). **(15)** Rain precipitation (mm). **(16)**Wind direction. **(17)** Wind Speed per meter (m/s).

Particulate Matter 2.5, which implies particles noticeable all around that are ≤2.5 microns in width. These particles are so little they can enter human bloodstream and cause all way of afflictions. PM2.5 levels are estimated in micrograms per cubic meter (1.0 × 10^-9 kg/m³ or 1 µg/m³). To place this in context, 50 micrograms (µg) weighs about as much as a unique finger impression.

Table 2: Air Calibre Indicators

$PM_{2.5}$	Air Calibre Indicator	$PM_{2.5}$ Health Effects
0 to 12.0	**Righteous 0 - 50**	Less risk or nil.
12.1 to 35.4	**Modest 51 - 100**	Respiratory issues might arise in one with poor immune system.
35.5 to 55.4	**Detrimental for Vulnerable persons 101 - 150**	Might result in respiratory related problems, might negatively effect the smooth functioning of lungs and heart. Also people with major ailments and elderly might get effected severely.
55.5 to 150.4	**Injurious 151 to 200**	Elevation in lung diseases and heart related issues. This level has a ability to prepone mortality in a man with cardiopulmonary malfunctioning. Average healthier people might start getting affected showing initial symptoms.
150.5 to 250.4	**Poisonous 201 to 300**	It will lead the previous stage to more dangerous aggressively with effects on lung and heart related disease is very much significant. Elevation in health issues with normal people.
250.5 to 500.4	**Hazardous 301 to 500**	This amount of contamination will lead to hazardous situation by presenting worst situation where people life is on stake. The pace at which the above mentioned health issues will increase and harm one will be very high.

Table 2 contains air calibre indicators based on PM2.5 value, sourced by Environmental Protection Agency.

V. IMPLEMENTATION

This segment talks about an execution of the four forecast strategies utilized for examination. Principle commitments of this paper are: (i) The usage of various supervised learning models as to estimate PM2.5 based on considered dataset. (ii) Comparison of regression models with respect to R2 Score. Fig 1 lists stages of implementation. This whole work is carried using Language Python and its libraries with Jupyter Notebook.

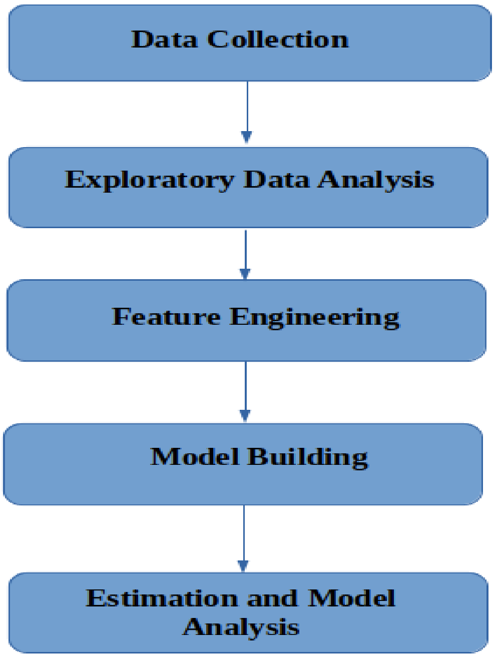


Fig 1: Flowchart of Implementation Stages

A. Exploratory Data Analysis

After loading of data exploratory data analysis second step to follow is Exploratory data analysis. After numerous long stretches of investigating and picturing the information it was apparent to decide the accompanying:
(i) Dealing with missing values (ii) Discovering Outlier and efficient handling of same. (iii) Correlation between PM2.5 and other parameters.

Around 6% of missing values were present in data in hand. Experiment was carried out with two set of values. By dropping the missing values and by strategically filling the missing values. Forward filling and mean filling was carried out and performance evaluation was done. By comparing the performance, model with dropped NA performed better than strategically filled values model. In line with model performance dropping was more appropriate than filling in missing values also percentage of missing values are considerably low. Dropping nan spared 32681 records for further execution.

Outliers may adversely impact the model performance. There present some records with PM2.5 value above 600 which could be considered as an outlier, as air monitoring gadget has measuring limit up till 500.These values were simply replaced by maximum value possible i.e. 500. Other than that there were no significant outliers in the information which was obvious from box plot as appeared underneath Fig 2.

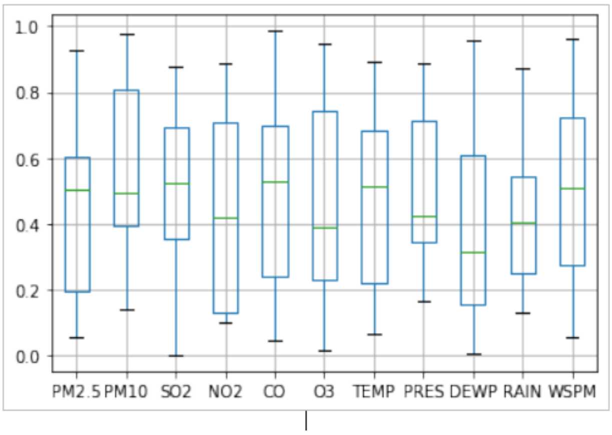
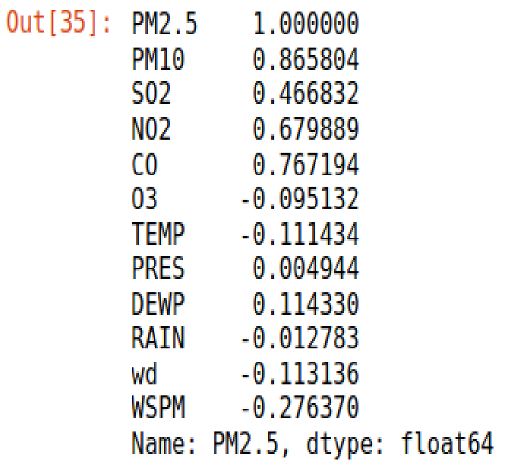


Fig 2: Box plot for Outlier Detection

B. Feature Engineering

This is a section where relevant features has to be selected for model building. The third step which is very much essential for best working model. The process of selecting the right feature is quite challenging since it requires analytical reasons as well as strong background knowledge about the respective area of study. Here is intact knowledge about various chemical compounds identified as pollutants,the factors which generate them and the cause they are going to have on biological life on earth. Some analysis is carried out as a part of this work where mainly correlation between each attribute is checked. Not only that,but correlation was exclusively checked against PM2.5 and other factors.

Since PM2.5 along with diverse chemical compounds exists during mutual transformation it's important that we dissect connection amongst PM2.5 and rest of air poisoning elements ahead of time. Fig 3 gives correlation values between PM2.5 and other attributes. Accepting all preparation information as tests, the impact of air pollutant factor on PM2.5 is also shown Figure 4. Figure shows the multiple scatter plots describing correlation amongst PM2.5 and other air contamination. The correlation among parameters plays major role in studying behaviors of parameters. Also selecting or deselecting parameters for further processing.



```
Out[35]: PM2.5    1.000000
         PM10     0.865804
         SO2      0.466832
         NO2      0.679889
         CO       0.767194
         O3      -0.095132
         TEMP    -0.111434
         PRES     0.004944
         DEWP     0.114330
         RAIN    -0.012783
         wd      -0.113136
         WSPM    -0.276370
         Name: PM2.5, dtype: float64
```

Fig 3 : Correlation among PM2.5 and Other Attributes

Fig 4 shows a pair plot which carries information about correlation between each pair of parameters. It very well may be observed that the convergence of PM2.5 has a specific level of positive connection with PM10, CO, NO2 and SO2 (0.865804,0.767194,0.679889,0.466832). As and when these pollutants value increases same will lead into increase in values of PM2.5 .Among all PM10 ha highest correlation with PM2.5. Since in light of the fact that there takes place physical and chemical transformation process among PM2.5 and various poisons,, invariably PM2.5 and PM10 has high probability of mutual transformation. Also, connection amongst PM2.5 and PRES is most minimal 0.004944.

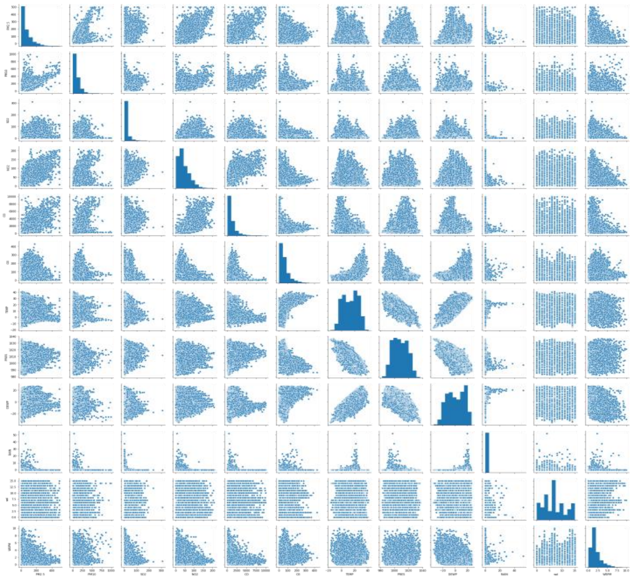


Fig 4: Correlation Matrix

Retrieval Number: F9912038620/2020©BEIESP
DOI:10.35940/ijrte.F9912.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

There is an easier way to visualize the correlation between factors i.e Heat Map elaborates information with Color Codes. Fig 5 is dedicated to present it here. Darker the color more is dependency.

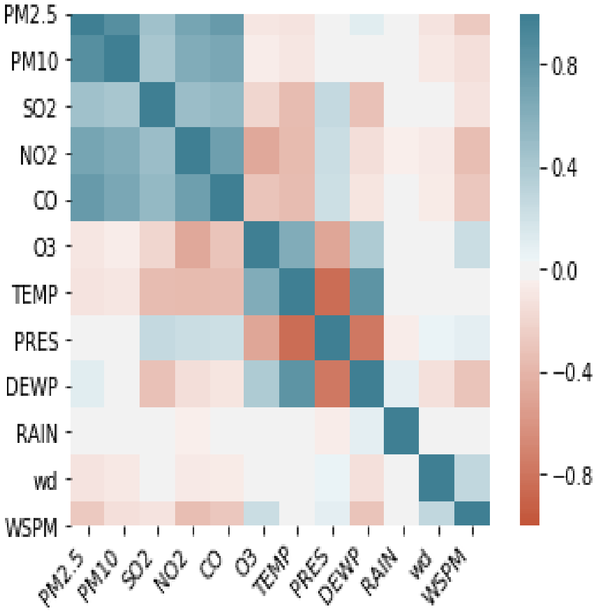


Fig 5: Heat map for relationship among attributes

The correlation coefficient between PM2.5 concentration and temperature, rain, O3, wind direction and wind speed per meter are negative with values -0.111434,-0.012783,-.0.095132,-0.113136 and -0.276370.That indicate that increase in value of above factors results in decrease in PM2.5. In many of the study it is observed fact that cooler the environment more the air contamination.

There was one analysis carried in order to get most effecting attributes with respect to PM2.5. sklearn ExtraTreesRegressor was used to identify most impacting features which resulted in graph below shown in Fig 6. Looking at this graph it is very much clear that PM10,CO,N02 are the major factors affecting the concentration of PM2.5.

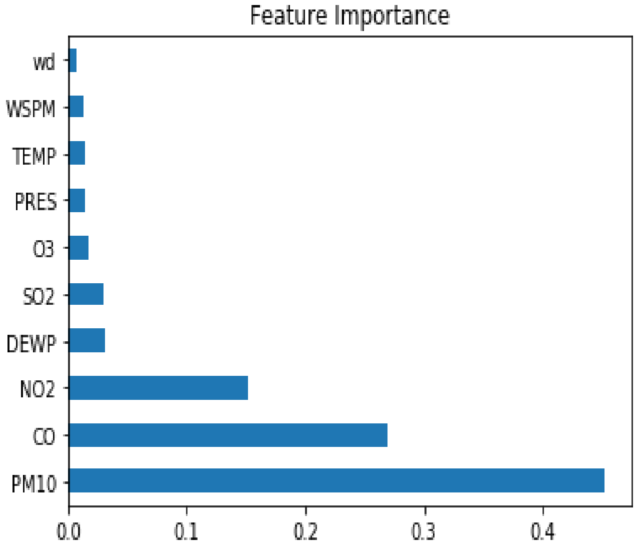


Fig 6: Feature Importance Indicator Graph

However current work is done by considering all the features but rain since it's evident from above plot. Figure 6 has no listing of attribute rain indicating zero dependency towards f6ature of interest.

Preprocessing

Preprocessing is the stage where data values are transformed into required format before feeding it to the algorithm. Data might contain attributes with mixed scale. Most of the learning models expects scaled values which helps them in being more efficient.

Scaling

It fundamentally assists with normalizing the information inside a specific range. Hence it likewise helps in accelerating the speed of algorithms. Scaling of data as also performed as to check the impact on model behavior. sklearn processing module gives an utility class StandardScalar which instrument Transformer API to in order to reckon average and Standard Deviation on preparation set in order to have the option to later reapply a similar change on the testing set.

Label Encoding

Vast majority of algorithms look forward numerical values and works only with numerical elements. Major requirement is transformation of categorical to numerical value. Sklearn gives an extremely productive apparatus to encoding categories into numeric values. It encode elements measurements between 0 and n_categories-1. In this current data set wind speed had six categorical values listed as E,EW,S,SE etc. Using LableEncoder it's been transformed to numerical.

C. Model Building

The fourth step of the process is Building a Model for precipitation of PM2.5. The following models were considered for predictive analysis of PM2.5. Those are Extreme Gradient Boosting Algorithm (XGBoost), Multiple Linear Regression (MLR), Random Forest(RF), and Multilayer Perceptron (MLP). Above mentioned four models were build with the intention of prediction of PM2.5 as well as to compare which algorithm does this job at it best. Also these models are cross verified with so called Base model i.e Mean model. This comparison will help us confirming that statistical models which are built are better than the Base/Mean Model. During the process of training it was made sure that same set of training and testing data are disposed to each models. Every model is pruned by setting its parameters at its optimum values as best as possible.

VI. RESULT ANALYSIS

Model evaluation is done by comparing each model performance with the base mean model. That shows how best the model is from mean model. Three following performance measuring factors meant for regression are considered here. i.e. MAE, RMSE R2 Score.

With following assumption that x1,…..,xn where n>=1 are values which are perceived for parameter x. The predicted values of y is given as y1,…,yn.

MAE is dictated as

$$MAE = 1/n \sum_{i=1}^{n} |xi - yi| \qquad (1)$$

Retrieval Number: F9912038620/2020©BEIESP
DOI:10.35940/ijrte.F9912.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

RMSE is dictated as

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (yi - xi)^2} \qquad (2)$$

R2 Score is dictated as

$$R2 = 1 - \frac{\sum (xi - yi)^2}{\sum (xi - \overline{y})^2} \qquad (3)$$

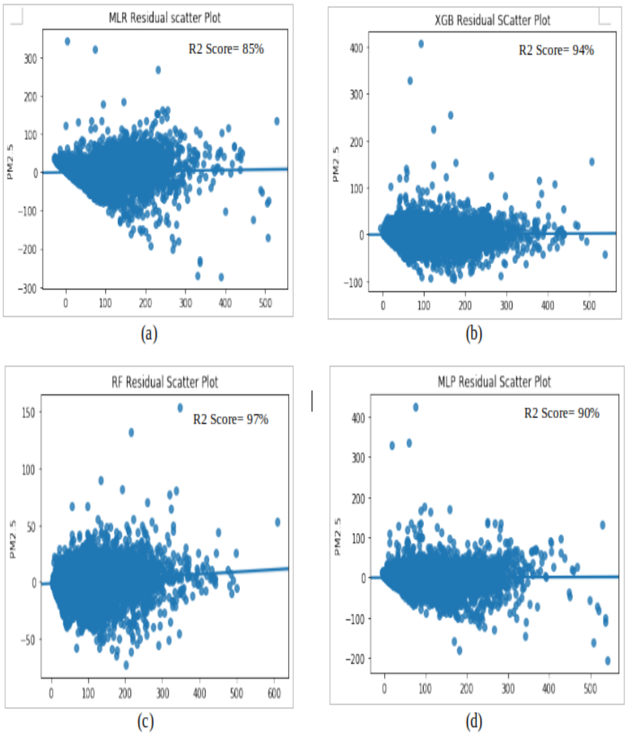
Residuals is an indicator of distance between data points and regression line. RMSE apprises, around the line of fit how distributed or concentrated the data is given in Eq1. Root mean square error is commonly used in prophesy or foretelling, closely associated with topography, regression task to verify experimental results. The mean absolute error is an average of the absolute errors given in Eq2. R-squared (R^2) or R2 Score shows statistic illustrating facts about the goodness of fit for a given model showed in Eq3. Table 3 describes the outcome of complete experiment with comparative tabular values. Result is quite visible and clear as how each model has performed with fed data.

Table 3: Tabulation of Model Performances along with Base Model

Models	MAE	RMSE	R2 SCORE
Base Mean Model (With scaling)	53.43	71.63	--
Multiple Linear Regressor (MLR)	18.28	27.43	85%
XGBoost Regression (XGB)	10.29	17.03	94%
Random Forest Regressor (RF)	7.68	11.34	97%
Multilayer Perceptron Regressor (MLP)	13.99	21.88	90%

It is clear from the above tabulation,Random Forest is pre-eminent model with MAE=7.68, RMSE=11.34, R2=97%. Very close competition for Random Forest is XGBoost with R2=94%, MLP and MLR with 90% and 85% respectively. Here if models are ranked then Random Forest Stands first followed by XGBoost in second place. Third is Multilayer Perceptron then Multiple Linear Regression. Former two have shown remarkably finest result among others for the estimation task in hand.

Fig 7 shows four residual scatter plots respectively for each model. Here each point represents one hour value where the prediction given by model is plotted across x-axis, where as accuracy of the prediction is plotted across y-axis. Residual is equal to perceived values minus estimated values. More the distance from line zero indicates bad the prediction is. The same can be viewed for this experiment in above figure. Fig 7(a) is a residual plot of MLR, where distance varies nearly from -100 to 100. Fig 7 (b) Indicates the variance from zero between -100 to 100 but compared to MLR points are more dense over line zero.Fig 7 (c) depicts the error variation plot of RF with variance being spread between -50 to 50 least in comparison with rest of the models. Finally Fig 7 (d) has residual plot with deviation between -100 to 100. Experimental result is evident to conclude that out of four models, the Random Forest method has emerged as an effective for PM2.5 concentration prediction.



(a)

(b)

(c)

(d)

Fig 7: Residual Scatter plot of a) MLR predictions b) XGB predictions c) RF predictions d) MLP predictions

VII. CONCLUSION AND FUTURE WORK

PM2.5, one of hazardous contamination causing serious health issues to all living being. All these effort is carried out with the intention that study might help the so society in at least tiniest. This study carried out by considering data belonging to specific city in Beijing. Reasonable amount of importance was given to Data exploration and Feature selection, as these both are essential steps in having a better models. Scaling of data was performed with Standard Scaling function. The four predictive models built were Multiple Linear Regressor, Random Forest Regressor, XGBoost Regressor and Multilayer Perceptron. Performnce is compared based on three mterics MAE, RMSE and R2 score. Result obtained suggests that Random Forest is foremost in predicting the PM2.5 with less error and with R2 Score of 97%, followed by which XGBoost algorithm with 94.5%, then Mutltilayer Perceptron with 90% and Multiple Linear Regression with 85% of R2 Score. Result is validated with plot of Residuals of four model prediction.

There is lot of scope for future work and enhancement. Data pruning can be applied more rigorously hence more precise prediction can be obtained. More ground work on environmental effects on concentration of PM2.5 can be performed by paying more attention and dedication to study how chemical reaction of various pollutants takes place. Meteorological features such as wind speed,wind direction effects can be studied. Not only prediction of PM2.5 but other dangerous pollutant prediction could be performed which are still a threat to human lives. Time series Analysis would take the study in different direction giving insights about environmental effects.

REFERENCES

1. David Núñez-Alonso, Luis Vicente Pérez-Arribas, Sadia Manzoor , and Jorge O. Cáceres Laser. February 2019. Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region. Hindawi Journal of Analytical Methods in Chemistry, Article ID 9753927,9 pages https://doi.org/10.1155/2019/9753927.
2. Bingyue Pan. 2018. ICAESEE 2017, Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conference Series: Earth and Environmental Science. Doi :10.1088/1755- 1315/113/1/012127.
3. Jihan Li,Xiaoli Li and Kang Wang. 15 October 2019, Atmospheric PM 2.5 Concentration Prediction Based on Time Series and Interactive Multiple Model Approach. Hindawi Advances in Meteorology Volume 2019, Article ID 1279565, 11 pages https://doi.org/10.1155/2019/1279565.
4. Mehdi Zamani Joharestani, Chunxiang Cao,Xiliang Ni, Barjeece Bashir, Somayeh Talebiesfandarani. 4 July 2019. PM 2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. Atmosphere 2019, 10, 373, www.mdpi.com/journal/atmosphere.
5. Simone Andréa Pozza.Ed Pinheiro,Tatiane Tagino Comin,Marcelino Luiz Gimenes,José Renato Coury, 2010. Time series analysis of PM2.5 and PM10-2.5 mass concentration in the city of Sao Carlos, Brazil. International Journal of Environment and Pollution.Vol. 41, Nos. 1/2, 2010.
6. Jan Kleine Deters, 1 Rasa Zalakeviciute, Mario Gonzalez, and Yves Rybarczyk.18 June 2017. Modeling PM 2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. Hindawi Journal of Electrical and Computer Engineering Volume 2017, Article ID 5106045, 14 pages https://doi.org/10.1155/2019/5106045.
7. Dixian Zhu ,Changjie Cai ,Tianbao Yang and Xun Zhou, 24 February 2018. A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big data and cognitive computing , MDPI.
8. Jiaming Shen, IEEE Honor Class. PM 2.5 concentration prediction using times series based data mining. Shanghai Jiao Tong University.
9. K. Mahesh Babu, J. Rene Beula,July 2019. Air Quality Prediction based on Supervised Machine Learning Methods. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-9S4.
10. Aditya C R, Chandana R Deshmukh , Nayana D K , Praveen Gandhi Vidyavastu.May 2018. Detection and Prediction of Air Pollution using Machine Learning Models. International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018.
11. Mahmoud Reza Delavar , Amin Gholami, Gholam Reza Shiran, Yousef Rashidi, Gholam Reza Nakhaeizadeh, Kurt Fedra and Smaeil Hatefi Afshar. 23 February 2019, A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran. International Journal of Geo-Information,ISPRS Int. J. Geo-Inf. 2019, 8, 99; doi:10.3390/ijgi8020099,MDPI.

AUTHORS PROFILE

Anusha Anchan working as an Assistant Professor in the Department of Computer Science & Engineering. She has published two research papers with International Publications. She has guided many undergraduate projects of which one got selected for FAER-McAfee Scholar Program. She has successfully completed IABAC Certified Data Scientist Course.

Manasa G.R. working as an Assistant Professor in the Department of Computer Science & Engineering. She has published one research article and attended several conferences. She has guided several undergraduate projects. She is a permanent member of MISTE.