The University of Tehran Press

Pollution

https://jpoll.ut.ac.ir/

Print ISSN: 2383-451X
Online ISSN: 2383-4501

Comparative Analysis of Machine Learning Models Vis-a-Vis Regularization Models on Predictive Ability of Pollution Levels in Bangalore

Vivekanand Venkataraman[1] | Mohan Babu G.N.[2]✉ | K.M. Sathish Kumar[1]

1.Department of Mechanical Engineering, BMS Institute of Technology & Management, Bengaluru – 54, Affiliated to Visvesvaraya Technological University Belagavi -18, Karnataka, India
2. AMC Engineering College, Bengaluru – 560083, Affiliated to Visvesvaraya Technological University Belagavi -18, Karnataka, India

Article Info

Article type:
Research Article

Article history:
Received: 18 April 2025
Revised: 2 August 2025
Accepted: 3 September 2025

Keywords:
Particulate Matter
Random Forest
Lasso
Ridge
Xgboost

ABSTRACT

The work brings in various dimensions to understand the importance of machine learning models in terms of predicting and understanding the variables which affect the concentration of Particulate matter $PM_{2.5}$ and PM_{10} for Bangalore city. In this work Metrological variables, Pollutants are considered as inputs. In addition, the work highlights the differences achieved in terms of performance metric especially in terms of error variance and prediction power for particulate matter using Regularization, Bagging and Boosting techniques. It specifically brings about areas where these techniques can perform well and underperform. The work also compares how the models performed with and without features such as seasons. It was noticed the order of feature importance differed for regularization, boosting and bagging models. It was noticed that Boosting techniques such as Xgboost had lower RMSE(9.1), MAPE (15.75) and higher R^2 values (.72) for $PM_{2.5}$ than other models however overfitting was noticed. However random forest had a lower R^2 (.64) compared to boosting and RMSE (10.35) and MAPE (22.45) were slightly larger and tendency to overfit was lower. To understand further, a new approach was created to diagnose where exactly these models perform well and underperform. The Absolute values were divided into percentile values and correlation was investigated with respect to error or residual values, adding to the uniqueness of this work. It was found that Extreme values tend to be correlated to larger residual values and those within the normal percentile range have lesser residual values.

Cite this article: Venkataraman, V., Mohan Babu G. N., & Sathish Kumar, K.M. (2025). Comparative Analysis of Machine Learning Models Vis-a-Vis Regularization Models on Predictive Ability of Pollution Levels in Bangalore. *Pollution*. 11(4), 1419-1432.
https://doi.org/10.22059/poll.2025.393694.2892

DOI: https://doi.org/10.22059/poll.2025.393694.2892

INTRODUCTION

The quality of air we breathe determines the quality of health we enjoy. Thus having good control over air quality is crucial. This work focusses on understanding the factors which affect particulate matter $PM_{2.5}$ and PM_{10} using Regularization and Machine Learning (ML) models. It also provides a novel approach in terms of analysis of understanding where ML models work well and fail in terms of performance metric/prediction. To accomplish this Data Cleaning, Exploratory Data Analysis and Feature Engineering are conducted based on which regularization and ML models are developed and compared. To understand where these models perform well and fail, correlation analysis was conducted between actual and residual values for various percentile range and accordingly conclusion were drawn, the above of segmenting and analysis is unique to this research methodology.

*Corresponding Author Email: *vivek999hyderabad@gmail.com*

The literature review in terms of understanding pollutants is varied and involve modelling using time series, machine learning, and neural networks. Monitoring and understanding of particulate matter started quite late as compared to other pollutants in fact monitoring of $PM_{2.5}$ started during 1997 (Department of Health and Human Services, 1997). In India, researchers have applied time series models such as AR, MA, and ARIMA modelling ((Sharma et al., 2009; Kumar & Jain.2010; Venkataraman et al.,2020) to estimate various pollutants and found non-stationary behaviour in pollutant levels. However when modelling as a whole and understanding cause and effect relationship it would be suitable to using machine learning models. Machine learning encompasses various mathematical models which include regression, and classification models (James et al., 2023). Simple to complex regression models (Polynomial, exponential) have been applied for pollutants such as $PM_{2.5}$, SO_2. Ozone levels were predicted using Classification and Regression Tree (CART) models by using two different sampling methods and the results obtained were similar (Bruno et al., 2004). Boosting regression tree were used to predict and obtain factors affecting NOx emitted due to jets (Carslaw & Taylor,2009). It was found that temperature and windspeed are significant factors Regression models have been applied to pollutants such as NOx, SO_2, and Total Particulate Suspended Matter (TSPM) by (Banerjee et al.,2011). (Napi et al.,2020) had used PCA to reduce multicollinearity and accordingly apply MLR and principal component regression for prediction of Ozone and it was found PCR provides better performance. It is noticed that most of the ML models were used in predicting Ozone, Nox and SO_2 pollutants. Only after a certain time where these models used for understanding particulate matter mainly because monitoring system for particulate matter were developed later. (Russo et al.,2015) used models such as stepwise regression for ranking followed by ANN to predict PM_{10} using previous PM_{10}, pollutants and metrological variables and the models showed less difference with other regression models. (Cortina–Januchs et al.,2015) also used clustering and MultiLayer Perceptron Neural Network to predict Daily average of PM_{10} and obtained Low MAE and MSE values and R^2 value between .67 and .78 for three stations. (Suleiman et al., 2016) used Various combinations of models by combing ANN with Principal component analysis(PCA), regularization models and showed that by including background poulltants better results were obtained. (Grange et al.,2018) predicted Daily average of $PM_{2.5}$ based on metrological variables and weather pattern data using Random Forest (RF) for evaluating 31 sites and R^2 varied between 54 to 71%. Mean Square Error ranged between 26 to 174. (Pan, B.,2018) found that Xgboost performs better than SVM, RF for prediction of Daily $PM_{2.5}$ using pollutant variables. (Jia et al.,2019) used Back Propagation Neural Network(BPNN) for predicting $PM_{2.5}$ for next hour with input as Metrological Variables, $PM_{2.5}$ (Previous 24 hour). (Chen et al., 2019) used regularization, bagging, and deep learning techniques to predict $PM_{2.5}$, NO_2 with land, road features and Satellite and Dispersion model estimates of target variables. They concluded more than one model should be used for recommendation. (Sihag et al.,2019) used various ML models to predict $PM_{2.5}$ with input variables pollutants and metrological variables. It was found RF was more suitable. (Ma et al.,2020) used deep learning techniques to predict $PM_{2.5}$ using past $PM_{2.5}$ data, metrological variables, and other highly correlated variables, it was found LSTM, CNN-LSTM had sound performance. (Chang et al.,2020) used Aggregated LSTM to predict hourly $PM_{2.5}$ and found ALSTM performs better than LSTM, SVM. (Gupta et al.,2021) used RF for predicting $PM_{2.5}$(Surface for Daily and hourly) and found Low RMSE and High R^2, however for large values there is quite amount of variation. (Kim et al.,2022) used bagging, boosting models for predicting hourly $PM_{2.5}$, PM_{10} using metrological variables, day, week, visibility and found that models tend to do overfitting. (Wang et al.,2023) used bagging, boosting models for predicting $PM_{2.5}$ with Input Variables Pollutants such asPM_{10},CO,NO_2,SO_2,O_3 and found CATBoost performs better. (Barthwal et al., 2023) used MLR, boosting, bagging, SVM models to predict $PM_{2.5}$, PM_{10} and found Gradient boosting and RF perform better than other models.

The usage of ML models such as RF, Support Vector Machines (SVM) have found varied applications such as estimating chlorophyll concentration in lakes (Sarvani etal.,2025). Likewise in terms of estimating water quality indices these algorithms and Xgboost has found to produce sound results thereby resulting in integrating with Web based geographic system (Sarvani etal.,2024). It is seen from the literature's that prediction have been done either based on daily average, hourly basis or 8hr basis of PM levels. In terms of inputs some papers incorporate past values of $PM_{2.5}$, PM_{10} mainly to predict future values in terms of hourly basis, this would increase R^2 and lower RMSE. In terms of performance metrics complex learning models such as random forest, Xgboost, Catboost, ANN, LSTM show similar performance but also tend to overfit. Based on the literature it was noticed that most of the models provided prediction based on hourly basis, however few of them addressed based on daily average. Addressing in terms of daily average and providing a generic information in terms of cause and effect would give a larger understanding of the mechanism of the relation between $PM_{2.5}$, PM_{10} and other variables. In addition the literature reviews have not provided where these models underperform in terms of performance metrics. This is a crucial issue as answers to this question can lead to further enhancement of ML models.

The current work has multiple objectives which include understanding the differences in terms of prediction for Regularization, Bagging and Boosting Models. To bring about the variables (pollutants and metrological factors) which affect daily average $PM_{2.5}$, PM_{10} in Bangalore city (Silk board Station). To analyse and present statistical evidence highlighting areas of strong and weak model performance. Bangalore in past few years have seen a rapid increase in terms of development activities leading to more amount pollution. Minimal research has been conducted in terms of understanding mechanism and modelling of pollutants for Bangalore city. Bangalore which is approximately 3000 feet above sea level can have severe winter and understanding the complex relationship of geography, pollutants is essential.

MATERIALS AND METHODS

Figure 1 provides a detailed description of the methodology in terms of system process which is Input, Transformation process and output.

Input Variables / Data collection

Silkboard region in Bangalore experiences huge amount of traffic congestion due to its proximity to IT hub, colleges, and universities. Data is collected in this region by Central Pollution Control Board (CPCB, https://airquality.cpcb.gov.in/). CPCB collects data is for various locations in Bangalore on a 15-minute frequency, in addition data set is available either in an hour, 8 hour or a day format, typically an hour format would imply averaging the 15 min data and a day format would be averaging the 15 min data set over a day. In the current study the data analysed was for daily average for various pollutants and metrological factors. The data collected was between 01-01-2019 to 31-12-2023 resulting in 1826 data points. The Input Process provides target variables and predictor variables for modelling purposes.

Data Cleaning (Data Preprocessing)

The data set is analysed using python, and various libraries such as matplotlib, sklearn, SciPy, statsmodel. In terms of extreme value there was a one value which was around 794 for $PM_{2.5}$ was removed and later imputed accordingly. Figure 2 provides the detail of the outlier, it is noticed that there is a missing value for PM_{10} and the other variables fall within the limits of typical occurrence of the values, hence it was decided this could have been an instrument error.

To conduct analysis missing values were investigated, it was found that the missing values are less than 10 percent, (Hair et al., 2013) mention's various thumb rules and highlights that

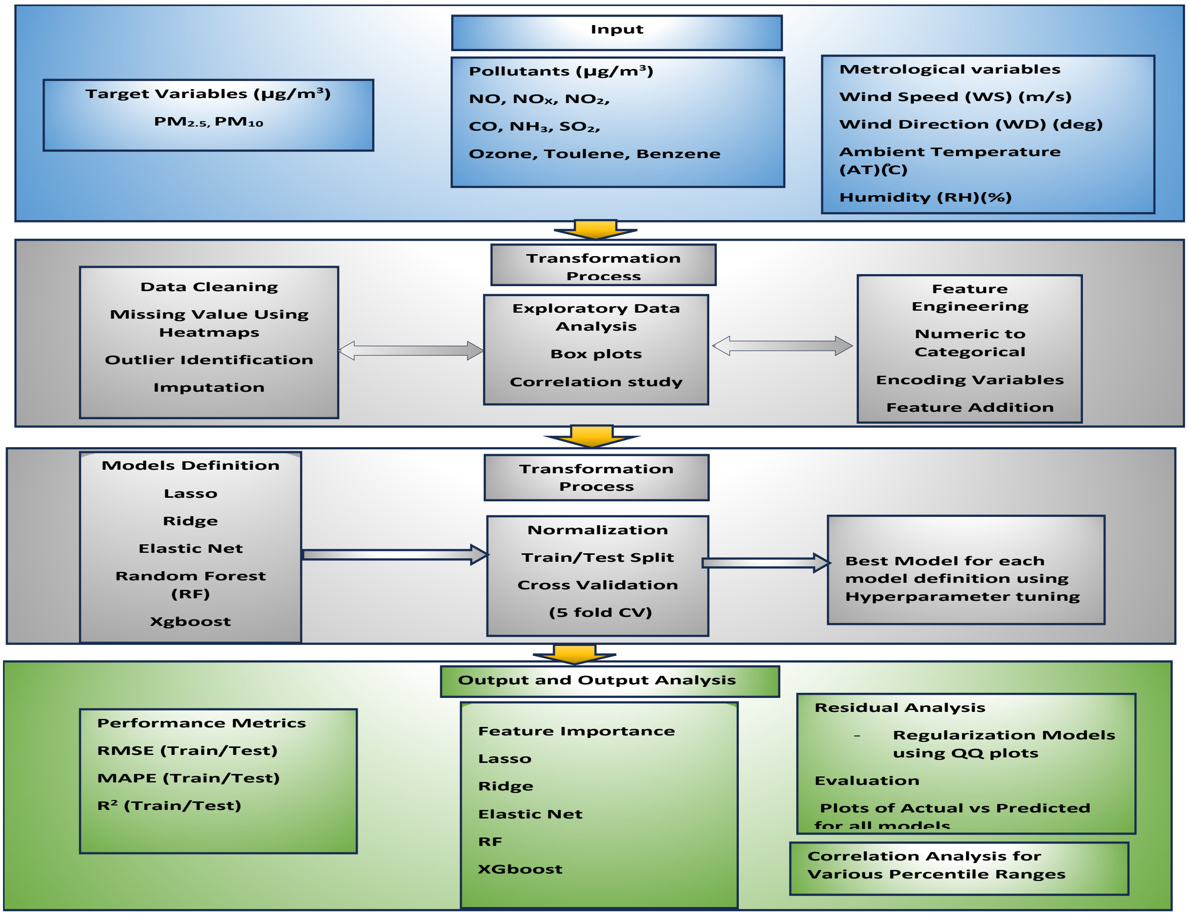
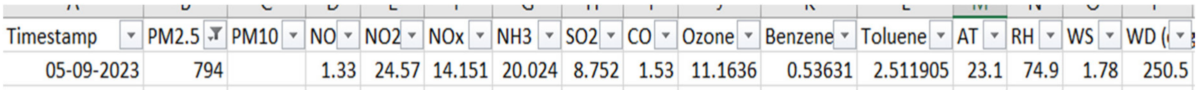


Fig. 1. Methodology



Timestamp	PM2.5	PM10	NO	NO2	NOx	NH3	SO2	CO	Ozone	Benzene	Toluene	AT	RH	WS	WD (
05-09-2023	794		1.33	24.57	14.151	20.024	8.752	1.53	11.1636	0.53631	2.511905	23.1	74.9	1.78	250.5

Fig. 2. Outlier for $PM_{2.5}$

for missing values less than 10 percent removing a variable is not advocated and any kind of imputation method can be used. It can be seen from the heat map (Figure 3) that most of missing values are random except for the variable NO and SO_2, in other cases a complete set of values in a row have missing values and the row missing values are random. It was found that the missing values for NO and SO_2 were during the month of July, August for the year 2020, hence monthly average was used for imputation.

Exploratory Data Analysis (*EDA*)
Seasonal patterns were observed for $PM_{2.5}$ and PM_{10} from the box plot in figure 4 with trend in terms of decreased values for summer and increased values for winter are observed. This could typically be due to temperature inversion wherein colder air would be at the bottom surface leading to lack of warm air. From the plot an inverse relationship between the meteorological variables and particulate matter when conditioned based on seasonality is noticed indicating seasonality as a factor. Based on Figure 5, some of the variables are highly correlated which include NOx with NO and NO_2 likewise NO and NO_2 , $PM_{2.5}$ and PM_{10}, also seen is mild correlation between $PM_{2.5}$ and humidity levels, Wind speed, Wind direction whereas PM_{10}

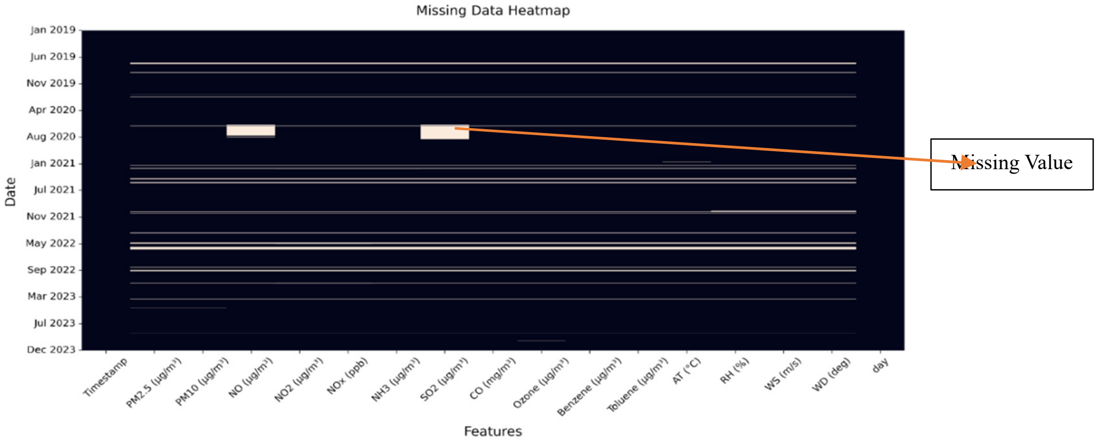


Fig. 3. Heat Map of Missing values for the Variables

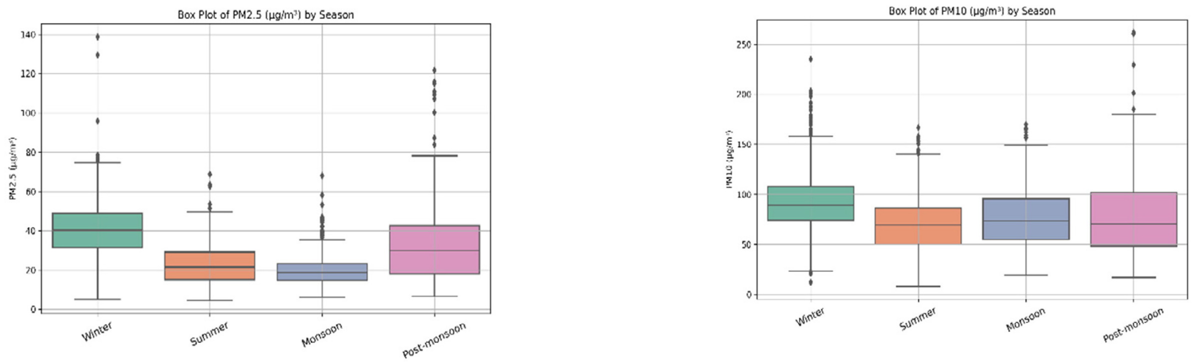


Fig. 4. Box Plot of Particulate Matter for Various Seasons

seems to be correlated with humidity levels. To understand specifically about the variation in the pattern, *seasonality as a feature is added* wherein winter includes months from December to March, summer from April to June, monsoon from July to September and Post monsoon October. It is noticed from plots that PM values are low and during winter season the values tend to be higher and reasons were explained before. Further exploration is conducted by observing changes in correlation when conditioned upon various season. It is noticed *from figure 5 that correlation increases for some variables when seasons are considered*, for example for the correlation of $PM_{2.5}$ with NOx across all years is .23 whereas when looked from each of the seasons the correlation shows an increase to .26 for summer and .31 for winter season, also seen is increase in negative relation with humidity (more than -.55) with post monsoon season close to -.66, in similar lines it can be said for PM_{10} and in some seasons the correlation value seems to be high close to .49 with respect to relation between pollutants and more than .5 with respect to metrological variables. Thus seasonality as a part of feature addition is created. After Addition of Holiday category the average values on holidays and on days taken a few days before were similar to non-holiday days, hence this category was not included.

Regularization and Machine Learning Models

Lasso, Ridge, Elastic Net, Random Forest Regressor, XGboost were evaluated using metrics such as Root Mean Square Error (RMSE), R square and Mean Absolute Percentage Error (MAPE). While running the models all the variables were standardized based on the usual

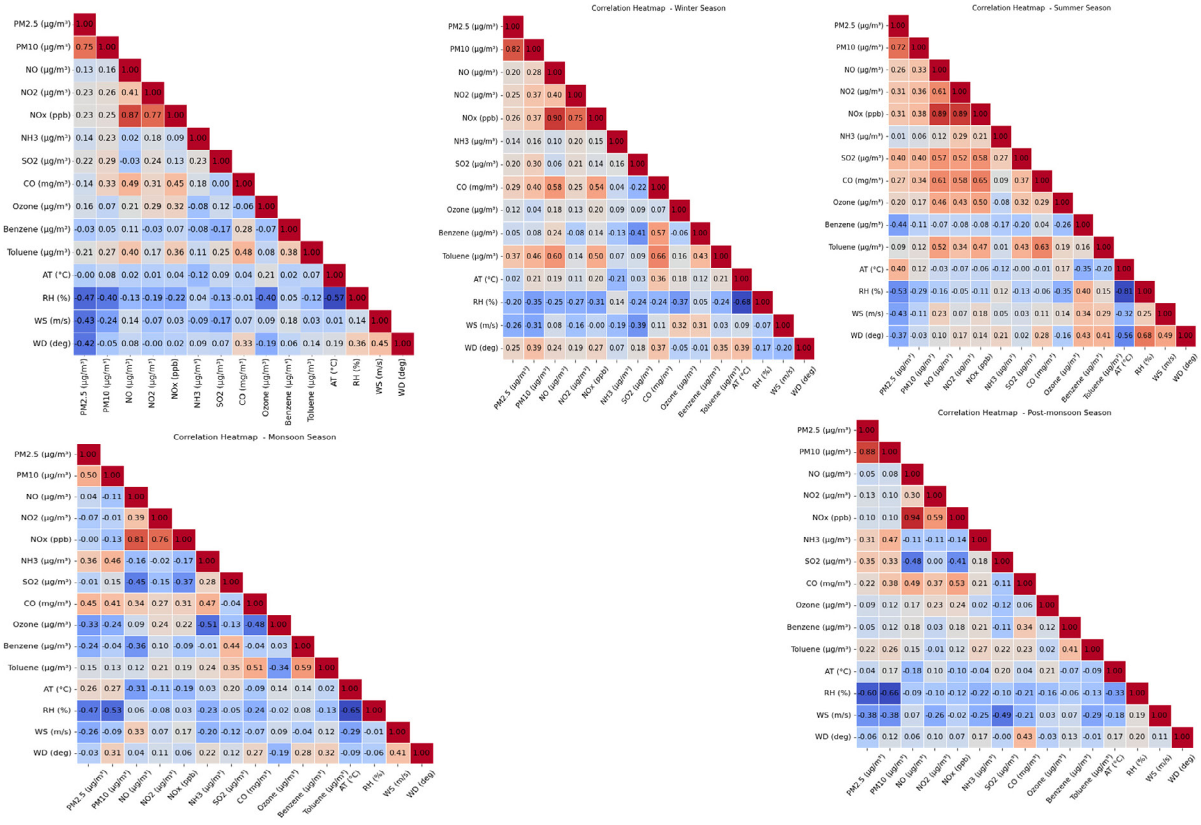


Fig. 5. Correlation for variables across all variables, for various seasons

standardization formula $z_i = \frac{X_i - \overline{X_i}}{S_i}$ *where* $\overline{X_i}$ *is mean,* S_i *standard deviation* .

Regularization models

Lasso, Ridge and Elastic net models

A detailed information on these models can be understood from (James et al., 2023). Based on this the models of these regularization technique have been described in mathematical terms. The concept of regularization is achieved by adding a penalty function to the ordinary least square (OLS) equation. The penalty function varies with the type of norm used which can be either L1 or L2 norm. In case of ridge L2 norm is used, in case of lasso L1 norm is used and for elastic net L1 and L2 norm are used. The OLS is given by $(y_i - \widehat{y_i})^2$ and in case of linear models $\widehat{y_i} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$ where β are the coefficients and x the observation value of the variables

Machine Learning Model Bagging & Boosting (Random Forest, XGBoost)

A detailed study on random forest is given by (Breiman, 2001). In random forest a set or forest of decision trees are formulated based on the principle of bagging and bootstrapping. Each decision tree is build based on using bootstrapped training data set with randomized set of features and accordingly a decision tree models is built likewise various other trees are built based on a similar mechanism of bootstrapping and selecting randomized variables for modelling. In Boosting the final model is developed based on set or ensemble of weak classifiers wherein weak classifiers imply a decision tree which can classify the output poorly. An ensemble of weak learners is to ensure that all weak leaners when combined would provide a sound model. There are different boosting techniques which include Adaboost, Gradient boosting and Xgboost. Gradient boosting in terms of mathematical modelling and algorithm has been explained by (James et al., 2023). and Xgboost by (Chen & Guestrin, 2016).

Train, Test data and Evaluation Metric

The models were run using 80 percent train data which and the selection of the train data set was randomized with five-fold cross validation of the train data set would be sound enough to estimate a good set of parameters and hyper parameters for the model. In this case grid search method was used to find out the parameters and hyper parameters. Typically for continuous data set three metrics are most often used which include RMSE, MAPE and R square.

RESULTS & DISCUSSION

Model Performance for $PM_{2.5}$ and PM_{10}

The regularization models, RF, Xgboost is run based on the train data for $PM_{2.5}$, PM_{10} and accordingly alpha, hyperparameters is tuned to minimize overfitting. Figure 6 provides the performance metric comparison, for the regularization models the RMSE, MAPE of the test and train values are similar indicating no over or underfitting *and the predicted $R^2$for PM_{10} is .57 indicating a reasonable fit and .50 for PM_{10}*. For *RF it seen that metrics is better than the regularization model and prediction power is higher in comparison to other models for both* $PM_{2.5}$, PM_{10} . *For Xgboost the performance measures are better and in terms of R^2 value the prediction power is higher than the rest but there are overfitting issues*. In terms of overfitting Xgboost has the highest difference between the train and test. Figures 7,8,9,10 provides the feature importance for the regularization, bagging and boosting models. Based on the graph the most important variables affecting $PM_{2.5}$ include RH levels, WS, AT, Winter Season, CO, WD (South East), Toluene, SO_2 . For RF the important features affecting $PM_{2.5}$ are humidity levels, wind speed, Seasons(winter), Toluene, SO_2, AT and the remaining variables seems to contribute less. For Xgboost Winter Season, Post Monsoon Season, RH, WD(SW) are important. It is worthwhile to mention that feature importance in terms of regularization is based on the coefficient value indicating that the large absolute value indicates a larger increase or decrease (based on Coeff value) in the target variable for every unit increase in the dependent variable. Regularization models typically work on linear relationship but by adding penalty to highly collinear variables. As seen from figure 4 it is known that highest correlation with respect to $PM_{2.5}$ are RH, WS, WD and this is reflected in terms of feature importance. But this is with the exception of AT which is least correlated. It should be noted that temperature and humidity are correlated and typically regularization models should penalize for it however it is not seen in this case. The models capture the winter season because encoding was created highlighting the difference between no winter and winter as significant and can be evidenced by the box plots. In case of RF the importance is measured based on variance reduction or impunity gain. Since

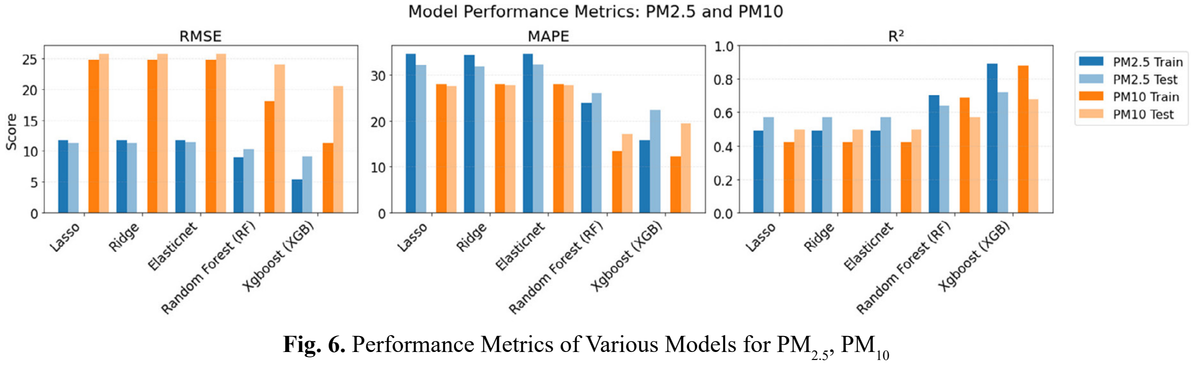


Fig. 6. Performance Metrics of Various Models for $PM_{2.5}$, PM_{10}

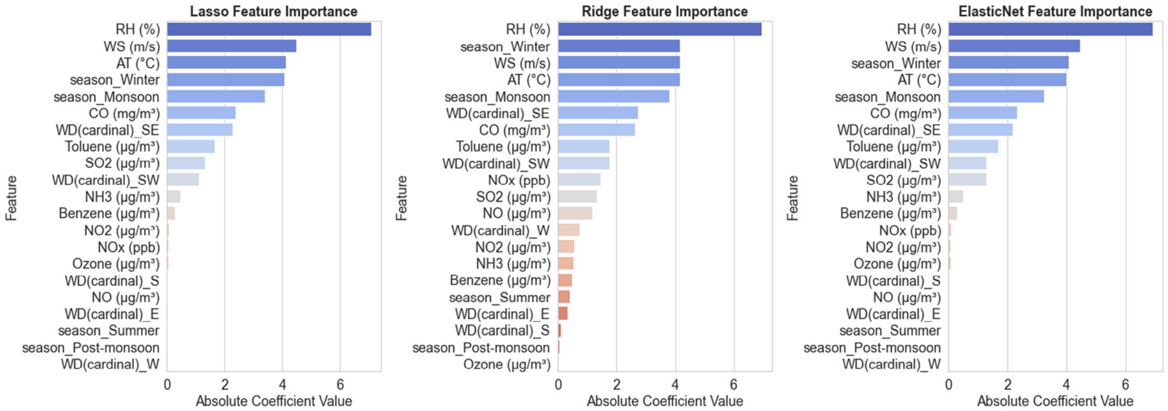


Fig. 7. Feature Importance for Regularization Models for $PM_{2.5}$

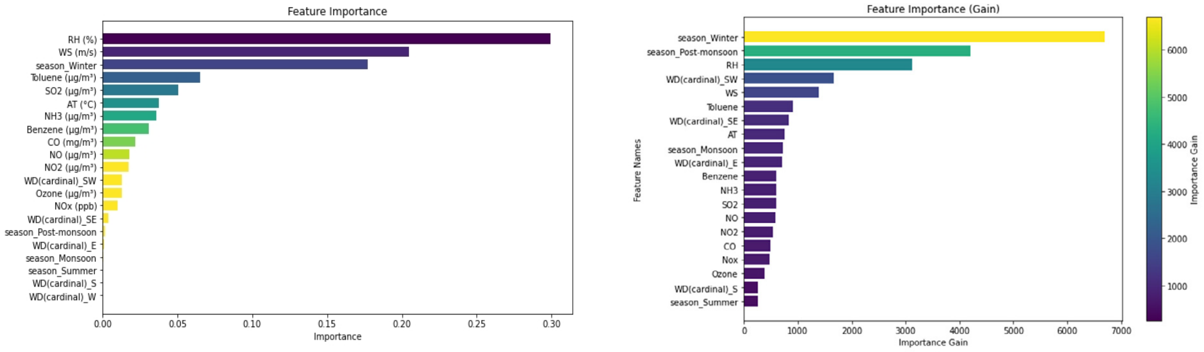


Fig. 8. Feature Importance for $PM_{2.5}$ (Random Forest, Xgboost)

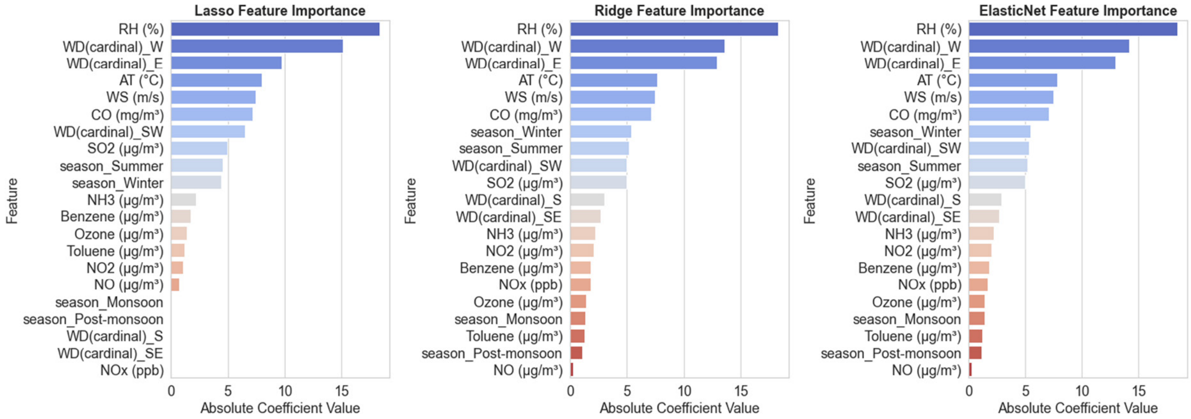


Fig. 9. Feature Importance for Regularization Models for PM_{10}

the output is continuous, variance reduction is computed, and whichever variable provides the highest variance reduction when the split occurs is considered an the most important variable. This can be thought of in similar terms of a regression model wherein removing a significant variable can lead to increase in Sum of Squares of Error or a reduction in Sum of Squares of Regression or in case of addition can lead to decrease in sum of square error term. In this model the interesting part is that continuous variables are captured as most important followed

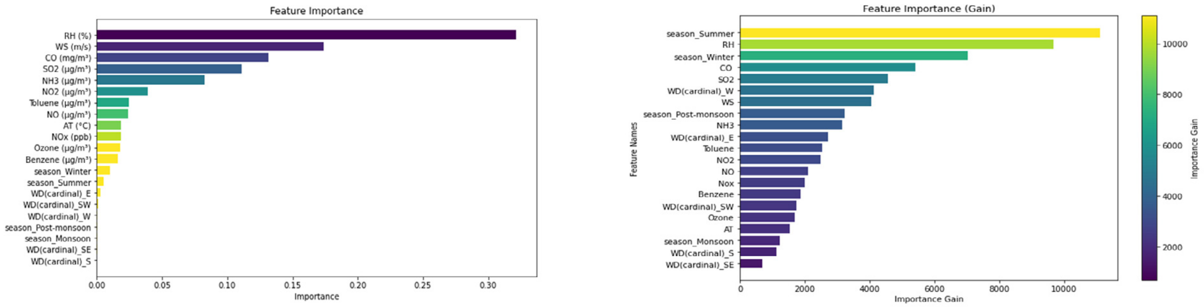


Fig. 10. Feature Importance for all models for PM_{10} (Random Forest, Xgboost)

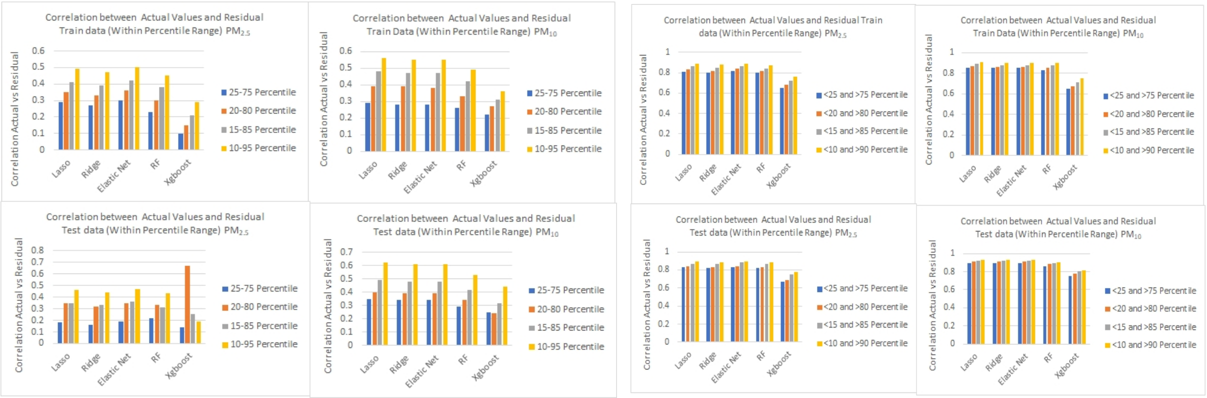
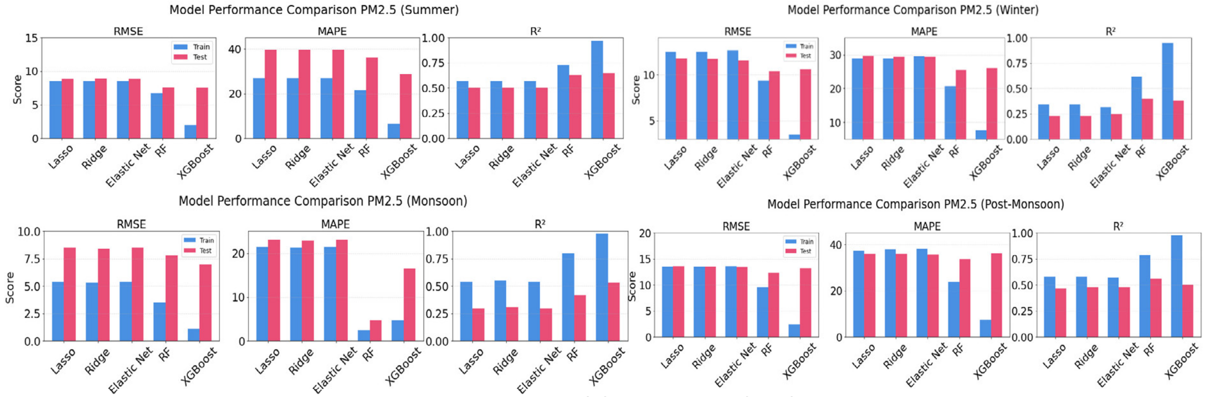


Fig. 11. Correlation between Actual Values and Residual (Outside Percentile Range) $PM_{2.5}$, PM_{10}

by categorical. There could be varied reasons for this one of them could be that there could be more splits for a continuous variable and this becomes significant when bootstrapping is used whereas when categorization is limited the splits can be minimal. Interestingly RF model captures toluene and SO_2 as important features, it should be noted that the correlation across all seasons is quite less but increases when conditioned upon summer and winter. This probably could be the reason why toluene is given importance may be split for winter season does create a condition for toluene and this may occur in multiple trees and further split from toluene may create significant reduction in variance. Xgboost considers gain based on the residue unlike random forest which uses variance, hence sensitivity in terms of importance would differ. In addition for Xgboost there seems to be more of categorical variable at the top than continuous variable. In this case it is winter season and post monsoon season. As seen in figure 4 boxplot there is change in terms of mean values when conditioned on the overall mean, since residue can capture small changes quickly the sensitivity in these changes can get reflected in terms of reduction and when there is a large change as shown in the plot the chances of winter season occurring as first leaf is higher. In case of continuous variable the splits become lot more which can lead to smaller changes in residue and may get captured at a later point.

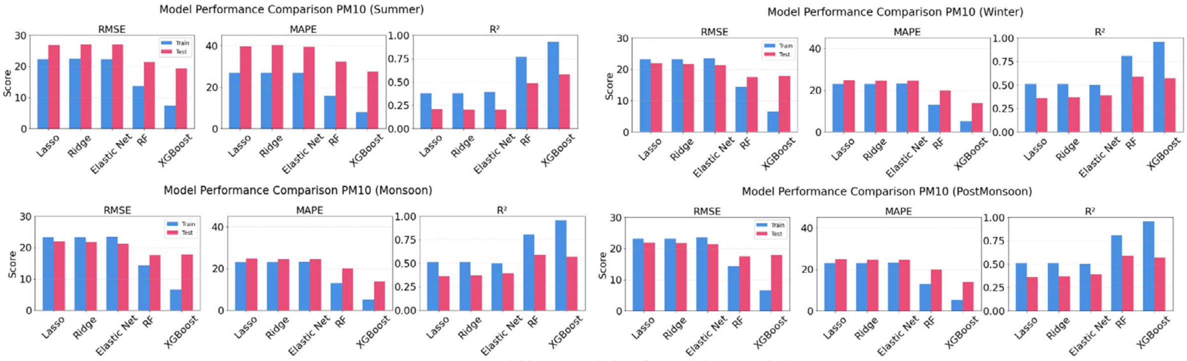
Model Prediction Power for $PM_{2.5}$ and PM_{10} within and outside various percentile

It should be noted that all the models do not provide very high values of R^2 and thus a need to investigate where the model is unable to predict accurately. *The purpose was to understand whether the large absolute value of residuals in terms of prediction occur due to large absolute values of actual values. The analysis was divided into various percentile values and based on the range of the percentile values the correlation between the actual values within these ranges were studied with respect to residual values for that range*. Figures 11 provides the various



Top 5 Important Features & Correlations Across Seasons (PM2.5)

Model	Summer	Winter	Monsoon	Post-Monsoon
Lasso	RH>WS>SO2>BNZ>NO2	WD_W>AT>RH>WD_E>WS	RH>BNZ>WS>TO>CO	RH>BNZ>WS>TO>CO
Ridge	RH>WS>SO2>BNZ>NO2	WS>TO>AT>RH>CO	RH>BNZ>WS>NO2>TO	RH>BNZ>NO2>TO>CO
Elastic Net	RH>WS>SO2>BNZ>NO2	WS>AT>TO>RH>CO	RH>BNZ>WS>NO2>TO	RH>BNZ>WS>TO>CO
RF	BNZ>SO2>RH>WS>CO	TO>RH>CO>WS>SO2	RH>CO>BNZ>NH3>O3	RH>SO2>WS>AT>NH3
XGBoost	BNZ>SO2>WS>RH>NO2	TO>RH>WD_E>WS>CO	BNZ>RH>CO>NH3>NO	RH>SO2>WS>WD_SE>BNZ
Top 5 Correlation	RH>BNZ>WS>SO2>CO	TO>CO>WS/NOX>WD/NO2>RH/SO2	RH>NH3>CO>O3>AT	RH>WS>SO2>NH3>TO/CO



Top 5 Important Variables & Correlations for PM10 (Season-wise)

Model	Summer	Winter	Monsoon	Post-Monsoon
Lasso	WD_W>RH>WD_SW>SO2>AT	WD_W>RH>WD_E>TO>NO2	RH>NH3>CO>WD_W>NOx	RH>WD_E>SO2>WD_SW>NH3
Ridge	RH>SO2>WD_W>NO2>WD_SE	Toluene>RH>WS>CO>WD_E	WD_W>RH>CO>NH3>WD	RH>SO2>WD_E>NH3>WD_SW
Elastic Net	RH>WD_W>SO2>WD_SE>NO2	RH>TO>WS>WD_E>CO	RH>WD_W>CO>NH3>WD	RH>WD_E>SO2>WD_SW>NH3
RF	SO2>NH3>NO>NOx>NO2	RH>TO>WS>SO2>WD_E	NH3>RH>CO>NOx>AT	RH>WS>NH3>AT>SO2
XGBoost	SO2>NOx>NH3>NO>WD_SW	WD_E>TO>RH>WS>SO2	RH>NH3>NO>CO>NOx	RH>WD_E>WS>NH3>SO2
Top 5 Correlation	SO2>NOx>NO2>CO>NO	TO>WD>CO>(NOx or NO2)>RH	RH>NH3>CO>WD>AT	RH>NH3>(CO or WS)>SO2>Toluene

Fig. 12. Performance of Models for various seasons and Comparison of Feature Importance

percentile ranges, it is seen clearly those within 25th percentile and 75th percentile values seem to be less correlated with residual value and those which are not within these percentile values are having larger corelation values. With increase in range of percentile the correlation increases indicating the model becomes quite unsuitable for values outside a certain range and is true for all the models. In this models *seem to be weaker in terms* of *predicting the extreme value*. Thus the extreme values seem to be a major contributor to lower R^2value and higher RMSE value and the values within the 25th and 75th percentile values contribute to lower RMSE values, MAPE values. Thus it can be said that these models seem to be quite good in predicting around a mean or median range but can fail at the extremes. *This indicates that modelling the higher percentile range values are a bit more complex*. It should also be mentioned that Xgboost has a slightly lesser correlation in terms of both inside and outside percentile range but still unable to predict the outside quartile range correctly.

It's important to explore models that can handle extreme values. While ANN and LSTM are complex models, they may not be the best models for explaining cause-effect relationships and may struggle with extremes. Some, like Reyes et al. (2010), used generalized extreme value distributions based on quantile patterns to model exceedances.

Model Prediction for $PM_{2.5}$ and PM_{10} conditioned based on seasons

It is noticed from figure 12 that for *$PM_{2.5}$* for all seasons that *most of the models do overfitting with XGboost being the highest*. Among all the models, Xgboost, random forest tends to predict better. The order of variable Importance for the regularization models are similar whereas for RF, Xgboost there is change noticed. *It is seen that RF and XGboost is able to capture 4 out of the 5 top variables in terms of correlation whereas the regularization models can capture 2 out of the 5 top variables.* This being more evident for Winter and Post Monsoon Season. Similar behaviour is exhibited for PM_{10} however the order is shuffled for few of them . *It seems clear that Xgboost, RF can capture linear relationship and is able to predict with smaller error variance, but not sound in capturing extreme values*.

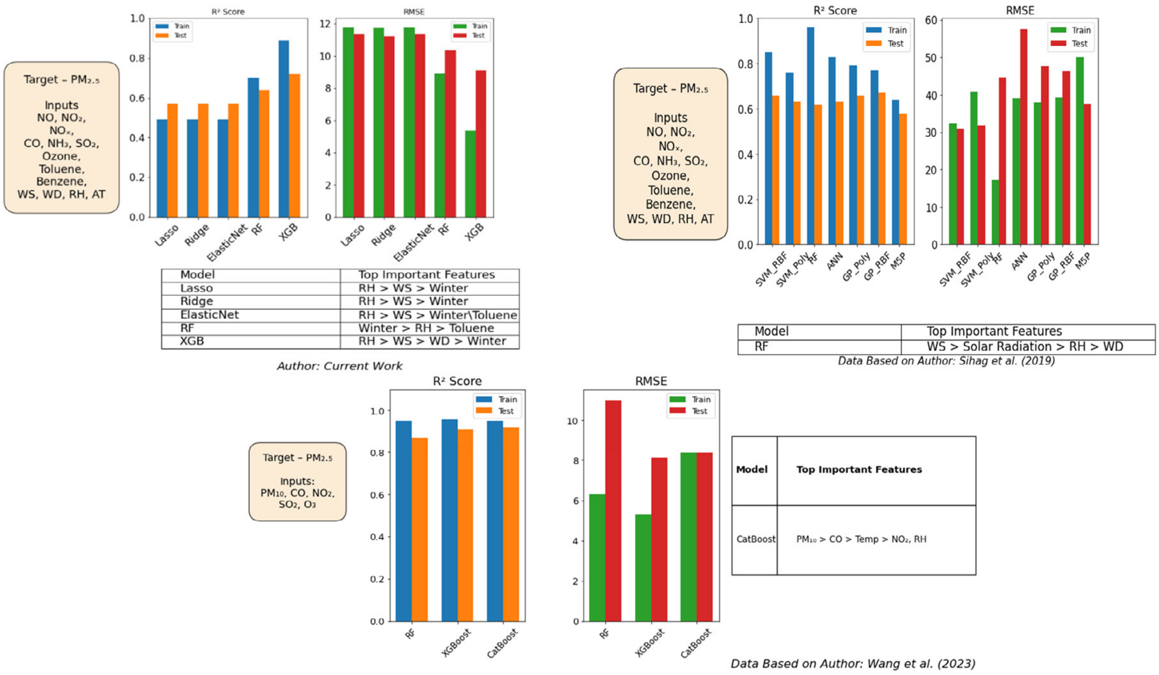
Regularization models show modest R^2 and RMSE values but exhibit low overfitting. In

contrast, Random Forest and XGBoost achieve higher R^2 and lower RMSE, with XGBoost showing a greater tendency to overfit. The important features identified vary significantly across regularization, bagging, and boosting models, complicating model selection. Therefore, it is essential to compare these results with existing studies, especially those that analyse model weaknesses and performance under extreme conditions. Figure 13 below provides the comparisons; it should be noted that data set and number of data points used in the models are different and the variables used as input differ. Based on Figure 13, it is seen there seems to be variation in terms of R^2 value and RMSE value. (Sihag etal.,2019) work has larger RMSE values and RF, ANN models tend to overfit. R^2 value also are in similar value to the current work. In terms of important variables RF gives importance to continuous variables. (Wang etal.,2023) provides CATBoost as the best model with sound RMSE and R^2 value however this is mainly due to inclusion of PM_{10} as Input variable. Thus it is seen that there are similarities with other papers in terms of overfitting and other performance metric. However these papers do not address the issues where they underperform especially with respect to extreme values.

Figure 14 provides comparison with respect to PM_{10} the current work has lower R^2 value and higher RMSE values when compared to (Suleima etal.,2016) the variables which added better performance was background pollutants, however the paper does not address how the model would work in case they are excluded. The paper has not addressed where the models can be weak in terms of extreme values. The current work addresses these issues in terms of quantiles.

CONCLUSION

The work highlights the differences achieved in terms of performance metric for particulate matter using Regularization, Bagging and Boosting techniques. This work also has identified the factors which play a role in affecting $PM_{2.5}$ and PM_{10} and highlights the issues in selecting a suitable model by providing a comparison of variation of feature importance across various model especially the choice of Random forest or Xgboost. When comparing with correlation analysis random forest produces similar results, but when box plots in terms of categories are chosen, Xgboost tallies well with feature importance. These issues were addressed before, but which model would be chosen for sound decision would require addition of more features. The models are weak in prediction with respect to extreme value, extreme values are quite important when it comes to providing alerts, it is a well-known fact that extreme events can cause severe damages. How do these models highlight this using known variables, is there a need of varied levels of data collection in terms of variables. Is there a need to gather information on traffic, construction activities and other activities. Is this possible in areas where development is



Model	Top Important Features
Lasso	RH > WS > Winter
Ridge	RH > WS > Winter
ElasticNet	RH > WS > Winter\Toluene
RF	Winter > RH > Toluene
XGB	RH > WS > WD > Winter

Author: Current Work

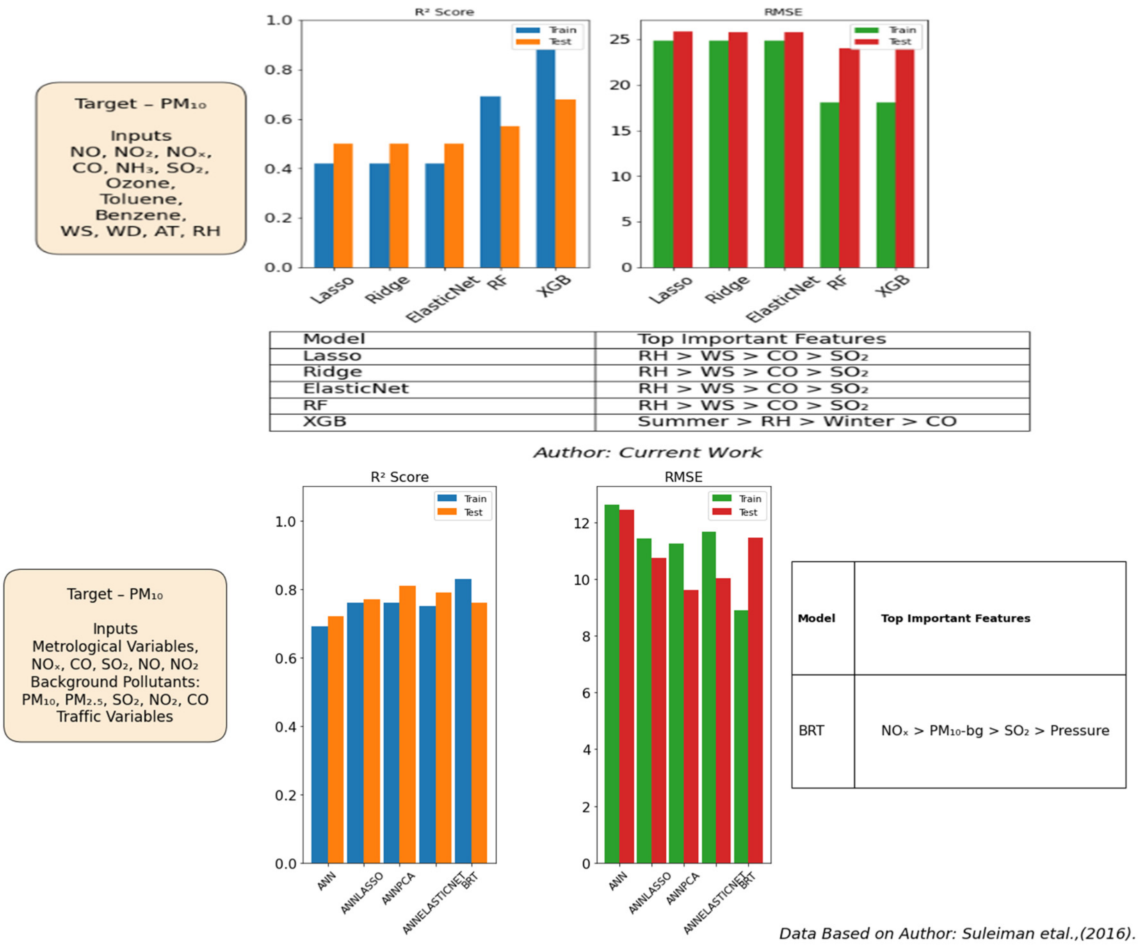
Model	Top Important Features
RF	WS > Solar Radiation > RH > WD

Data Based on Author: Sihag et al. (2019)

Model	Top Important Features
CatBoost	PM_{10} > CO > Temp > NO_2, RH

Data Based on Author: Wang et al. (2023)

Fig. 13. Comparison of Current Work with Sihag etal., 2019 and Wang etal., 2023 for $PM_{2.5}$



Model	Top Important Features
Lasso	RH > WS > CO > SO_2
Ridge	RH > WS > CO > SO_2
ElasticNet	RH > WS > CO > SO_2
RF	RH > WS > CO > SO_2
XGB	Summer > RH > Winter > CO

Author: Current Work

Model	Top Important Features
BRT	NO_x > PM_{10}-bg > SO_2 > Pressure

Data Based on Author: Suleiman etal.,(2016).

Fig. 14. Comparison of Current Work with Suleiman etal., 2016 for PM_{10}

happening, would there be investment in terms of money value, these are unanswered questions and these should lead to further progress in modelling. Can models predict with limited information if so to what extent are some issues which futuristic models should approach and, in such cases, the optimal choice should be provided.

GRANT SUPPORT DETAILS

NO: The present research did not receive any financial support.

CONFLICT OF INTEREST

NO: The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

NO: No life science threat was practiced in this research.

ACKNOWLEDGEMENT

We would like to thank the reviewers for providing suggestion which enhanced the quality of the manuscript.

REFERENCES

Banerjee, T., Singh, S. B., & Srivastava, R. K. (2011). Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India. Atmos. Res., 99(3–4), 505–517.

Barthwal, A., Acharya, D., & Lohani, D. (2023). Prediction and analysis of particulate matter (PM2.5 and PM10) concentrations using machine learning techniques. J. Ambient Intell. Humaniz. Comput., 14(3), 1323–1338.

Breiman, L. (2001). Random forests. Mach. Learn., 45, 5–32.

Bruno, F., Cocchi, D., & Trivisano, C. (2004). Forecasting daily high ozone concentrations by classification trees. Environmetrics, 15(2), 141–153.

Carslaw, D. C., & Taylor, P. J. (2009). Analysis of air pollution data at a mixed source location using boosted regression trees. Atmos. Environ., 43(22–23), 3563–3570.

Chang, Y. S., Chiao, H. T., Abimannan, S., Huang, Y. P., Tsai, Y. T., & Lin, K. M. (2020). An LSTM-based aggregated model for air pollution forecasting. Atmos. Pollut. Res., 11(8), 1451–1463.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., ... & Hoek, G. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int., 130, 104934.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (pp. 785–794).

Cortina–Januchs, M. G., Quintanilla–Dominguez, J., Vega–Corona, A., & Andina, D. (2015). Development of a model for forecasting of PM10 concentrations in Salamanca, Mexico. Atmos. Pollut. Res., 6(4), 626–634.

Department of Health and Human Services. (1997, July 18). Protection of human subjects; reports of the President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. Fed. Regist., 62(138), 38477–38480. https://www.govinfo.gov/content/pkg/FR-1997-07-18/pdf/97-18577.pdf

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM10 trend analysis. Atmos. Chem. Phys., 18(9), 6223–6239.

Gupta, P., Zhan, S., Mishra, V., Aekakkararungroj, A., Markert, A., Paibong, S., & Chishtie, F. (2021). Machine learning algorithm for estimating surface PM2.5 in Thailand. Aerosol Air Qual. Res., 21(11), 210105.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2013). Multivariate data analysis (8th ed.). Pearson Education Limited.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: With applications in Python. Springer International.

Jia, M., Cheng, X., Zhao, T., Yin, C., Zhang, X., Wu, X., ... & Zhang, R. (2019). Regional air quality forecast using a machine learning method and the WRF model over the Yangtze River Delta, East China. Aerosol Air Qual. Res., 19(7), 1602–1613.

Kim, B. Y., Lim, Y. K., & Cha, J. W. (2022). Short-term prediction of particulate matter (PM_{10} and $PM_{2.5}$) in Seoul, South Korea using tree-based machine learning algorithms. Atmos. Pollut. Res., 13(10), 101547.

Kumar, U., & Jain, V. K. (2010). ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO). Stoch. Environ. Res. Risk Assess., 24, 751–760.

Li, Y., Sha, Z., Tang, A., Goulding, K., & Liu, X. (2023). The application of machine learning to air pollution research: A bibliometric analysis. Ecotoxicol. Environ. Saf., 257, 114911.

Liaw, A. (2002). Classification and regression by randomForest. R News, 2(3).

Ma, J., Yu, Z., Qu, Y., Xu, J., & Cao, Y. (2020). Application of the XGBoost machine learning method in PM2.5 prediction: A case study of Shanghai. Aerosol Air Qual. Res., 20(1), 128–138.

Napi, N. N. L. M., Mohamed, M. S. N., Abdullah, S., Mansor, A. A., Ahmed, A. N., & Ismail, M. (2020, December). Multiple linear regression (MLR) and principal component regression (PCR) for ozone (O3) concentrations prediction. In IOP Conf. Ser.: Earth Environ. Sci. (Vol. 616, No. 1, p. 012004). IOP Publishing.

Pan, B. (2018, February). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. In IOP Conf. Ser.: Earth Environ. Sci. (Vol. 113, p. 012127). IOP Publishing.

Reyes, H. J., Vaquera, H., & Villaseñor, J. A. (2010). Estimation of trends in high urban ozone levels using the quantiles of GEV. Environmetrics, 21(5), 470–481.

Russo, A., Lind, P. G., Raischel, F., Trigo, R., & Mendes, M. (2015). Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmos. Pollut. Res., 6(3), 540–549.Saravani, M. J., Noori, R., Jun, C., Kim, D., Bateni, S. M., Kianmehr, P., & Woolway, R. I. (2025). Predicting Chlorophyll-a concentrations in the world's largest lakes using Kolmogorov–Arnold networks. Environ. Sci. Technol., 59(3), 1801–1810.

Saravani, M. J., Saadatpour, M., & Shahvaran, A. R. (2024). A web GIS based integrated water resources assessment tool for Javeh Reservoir. Expert Syst. Appl., 252, 124198.

Sharma, P., Chandra, A., & Kaushik, S. C. (2009). Forecasts using Box–Jenkins models for the ambient air quality data of Delhi City. Environ. Monit. Assess., 157, 105–112.

Sihag, P., Kumar, V., Afghan, F. R., Pandhiani, S. M., & Keshavarzi, A. (2019). Predictive modeling of PM2.5 using soft computing techniques: Case study—Faridabad, Haryana, India. Air Qual. Atmos. Health, 12(12), 1511–1520.

Suleiman, A., Tight, M. R., & Quinn, A. D. (2016). Hybrid neural networks and boosted regression tree models for predicting roadside particulate matter. Environ. Model. Assess., 21, 731–750.

Venkataraman, V., Prasad, S., Aswathanarayana, B., Barigidad, S., & Nayak, V. (2020). Development of time series models for various pollutants in Bangalore city using the Akaike information criterion. Eng. Appl. Sci. Res., 47(3), 249–263.

Wang, S., Ren, Y., & Xia, B. (2023). PM2.5 and O3 concentration estimation based on interpretable machine learning. Atmos. Pollut. Res., 14(9), 101866.